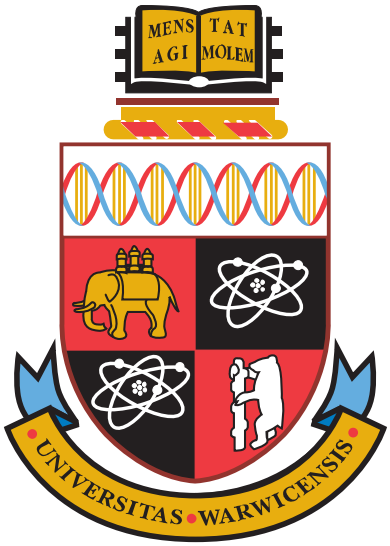


An Investigation into the Role and Mechanism of Action of Small Ubiquitin-like Modifier Interacting Motifs in *Arabidopsis thaliana* Proteins.

Stuart Nelis

Thesis

Submitted to the University of Warwick

for the degree of

Doctor of Philosophy

Systems Biology Doctoral Training Centre

September 2014

SYSTEMS BIOLOGY
DOCTORAL TRAINING CENTRE

Contents

List of Figures

List of Tables

Acknowledgements

I would like to give a special thank you to my two supervisors, Dr Ari Sadanandom and Dr Jay Moore for their guidance, support and patience throughout this project. I would like to thank Dr Ari Sadanandom for giving me the flexibility to explore novel research methods and for his support to make those approaches a reality. I would like to thank Dr Jay Moore for his help and support with the design of the computational aspects of this project.

Thank you to all the members of the Warwick Systems Biology Department who have helped me over the years with a special mention for Anne Maynard for helping sort out all things administration related, usually at the very last minute. I would like to thank the following people for their specific support: Professor George Baillie for making the peptide arrays, Professor Richard Napier for help with SPR, Dr Prashant Pyati at Durham University for help with *Pichia pastoris* and Dr Laura Baxter for supplying plant ortholog data.

Finally, I would like to thank my family, friends and my partner for believing in me and supporting me all the way through the Ph.D. project.

Declaration and Inclusion of Material from a Prior Thesis

This thesis is submitted to the University of Warwick in support of my application for the degree of Doctor of Philosophy. It has been composed by myself and has not been submitted in any previous application for any degree apart from the background material in the introductions of Chapters 4 and 5 which was previously submitted for a Master of Science in Systems Biology. The introductions to these two chapters refer to work described in a previous thesis that the plant DELLA protein GAI can be SUMOylated.

The work presented was carried out by the author except for the manufacture of large-scale cellulose peptide arrays described in Chapter 3. The arrays were synthesised by Professor G. Baillie's research group at Glasgow University as part of a collaboration.

The work described in Chapters 4 and 5 was published in the following two articles:

Conti, L. Nelis, S. Zhang, C. Woodcock, A. Swarup, R. Galbiati, M. Tonelli, C. Napier, R. Hedden, P. Bennett, M & Sadanandom, A. (2014) Small ubiquitin-like modifier protein SUMO enables plants to control growth independently of the phytohormone gibberellin. *Developmental Cell*, **28**(1): 102-10.

Nelis, S. Conti, L. Zhang, C. & Sadanandom, A. (2014) A functional Small Ubiquitin-like Modifier (SUMO) interacting motif (SIM) in the gibberellin hormone receptor GID1 is conserved in cereal crops and disrupting this motif does not abolish hormone dependency of the DELLA-GID1 interaction. *Plant Signaling & Behavior.* (Accepted September 2014)

Abstract

SUMO is a small protein that is ligated to other proteins to regulate their function. Ligation occurs at lysine residues within a SUMO site motif. A wide range of proteins are targets of SUMOylation and in plants SUMO plays a diverse role in many important processes. Processes including development, stress tolerance, hormone regulation, DNA repair and chromatin remodelling are regulated by SUMOylation. SUMO affects protein function primarily by establishing interactions through SUMO interacting motifs (SIMs) in interacting protein partners. SUMO can also alter protein function by blocking access to protein domains and by causing conformational changes to the target. The ability to predict SIMs in plant proteins would be useful for research into the poorly understood mechanisms behind SUMO regulation. Large arrays of synthetic peptides were screened with SUMO to identify SIM peptides. These data were used to characterise the sequence composition of plant SIMs. The plant SIMs were compared and contrasted with human SIMs to highlight the functional differences between these two evolutionary distinct species. The data were used to build a predictor for SIMs using random forest models. A new SUMO site predictor was built using random forest models as well. The SIM predictor was used to identify putative SIM containing proteins in the *Arabidopsis thaliana* genome and the functional enrichment of these genes was analysed. The role of SUMO in the plant gibberellin (GA) pathway was also investigated. The DELLA protein RGA is a negative regulator of GA signalling and this protein was shown to be SUMOylated. RGA stability is regulated by the GA receptor GID1 and it was demonstrated that GID1a contains a SIM. It was proposed that SUMOylated RGA interacted with GID1a through its SIM which inhibited its function. The model was tested by investigating the binding of SUMO to GID1a and by generating mutants of GID1a that had reduced SUMO affinity. The results demonstrate that GA signalling can be enhanced by introducing a mutation into the GID1a SIM.

Abbreviations

3-AT	3-amino-1,2,4-triazole
A. tumefaciens	*Agrobacterium tumefaciens*
AD	GAL4 activation domain
AUC	Area under ROC curve
BD	GAL4 DNA binding domain
BLAST	Basic local alignment search tool
bp	Base pairs
C-terminal	Carboxyl-terminal
CI	Confidence interval
Co-IP	Co-immunoprecipitation
DSB	Double strand break
E. coli	*Escherichia coli*
ECL	Enhanced chemiluminescence
EDTA	Ethylenediaminetetraacetic acid
FPLC	Fast protein liquid chromatography
FPR	False positive rate
GO	Gene ontology
GST	Glutathione *S*-transferase
GUI	Graphical user interface
HA	Influenza virus hemagglutinin protein

HPLC	High-performance liquid chromatography
HR	Homologous recombination
HRP	Horse radish peroxidase
HSDS	Homologous structure derived substitution
HSL	Hormone sensitive lipases
HSP	Heat shock protein
IMAC	Immobilised metal affinity chromatography
IPTG	Isopropyl β-D-1-thiogalactopyranoside
kDa	Kilodalton
LB	Luria-Bertani
MALDI TOF	Matrix-assisted laser desorption/ionization time of flight
MBP	Maltose binding protein
MS	Murashige and Skoog
MUSCLE	Multiple sequence comparison by log- expectation
N-terminal	Amino-terminal
Ni-NTA	Nickel-Nitrilotriacetic acid
NMR	Nuclear magnetic resonance
OD_{600}	Optical density at 600 nm
OOB	Out of bag
OTS	Overly tolerant to salt gene/protein
P. pastoris	*Pichia pastoris*
PAC	Paclobutrazol
PCA	Principal component analysis
PCR	Polymerase chain reaction
PIASx	Protein inhibitor of activated STAT x protein/gene
PTM	Post-translational modification

PVDF	Polyvinylidene fluoride
Rb	Rabbit
RBS	Ribosome binding site
RF	Random forest
ROC	Receiver operating characteristic
RU	Response units
S. cerevisiae	*Saccharomyces cerevisiae*
SAE	SUMO activating enzyme gene/protein
SA	Salicylic acid
SBP	SUMO binding protein
SCE	SUMO conjugating enzyme gene/protein
SDM	Site directed mutagenesis
SDS-PAGE	Sodium dodecyl sulphate polyacrylamide gel electrophoresis
SDS	Sodium dodecyl sulphate
SIM	SUMO interacting motif
SOB	Super optimal broth
SOC	Super optimal broth with catabolite repression
SPR	Surface plasmon resonance
Sp	Specificity
SUMO	Small ubiquitin-like modifier
SUM	SUMO protein/gene
TBST	Tris buffered saline with Tween 20
TEMED	Tetramethylethylenediamine
TE	Tris EDTA (buffer)
TIFF	Tagged image file format
UTR	Untranslated region

v/v	Volume/volume
w/v	Weight/volume
X-α-Gal	5-bromo-4-chloro-3-indolyl α-D-galactopyranoside
YPG	Yeast extract, peptone and glycerol

Amino acid abbreviations

Throughout this thesis, single letter abbreviations for the amino acids are used. At times the polarity and charge is mentioned. For clarity, the table below provides an explanation of these abbreviations and the chemical category that the amino acids belong to. To specify sequences with ambiguous positions, Perl style regular expressions are used. To specify any amino acid, both 'x' and '.' are used. Ψ is used to specify large hydrophobic amino acids. Square brackets are used to signify that a position can take any one of the amino acids within the brackets, e.g. [DE] indicates a single D or E. The caret character (^) inside square brackets indicates any amino acid *except* those inside the brackets, e.g. [^DE] indicates any amino acid except D or E.

Type	**Single letter code**	**Amino acid**
Small hydrophobic	G	Glycine
	A	Alanine
Large hydrophobic (Ψ)	V	Valine
	L	Leucine
	I	Isoleucine
	M	Methionine
	P	Proline
	F	Phenylalanine
	W	Tryptophan
Polar	S	Serine
	T	Threonine
	N	Asparagine
	Q	Glutamine
	C	Cysteine
	Y	Tyrosine
Positively charged	K	Lysine
	R	Arginine
	H	Histidine
Negatively charged	D	Aspartic acid
	E	Glutamic acid

Chapter 1

Introduction

Small ubiquitin-like modifier (SUMO) is a small protein that is conserved across eukaryotes (Iyer *et al.*, 2006) and plays a vital role in plant physiology (Hay, 2005). SUMO belongs to the group of ubiquitin-like proteins that all share the β-grasp fold structure of ubiquitin, the founding member of this group. The tertiary structure of SUMO is very similar to that of ubiquitin but the sequence similarity between the proteins is low which gives rise to very different biochemical properties (Burroughs *et al.*, 2007) and each protein plays a very different role in the cell.

Both ubiquitin and SUMO are ligated to other proteins as post translational modifications (PTMs) to regulate their function. The primary role of ubiquitin is to regulate protein stability through the formation of poly-ubiquitin chains. Proteins tagged with poly-ubiquitin chains are targeted for degradation by the 26S proteasome and ubiquitin dynamically regulates protein stability in response to various molecular cues (Hershko & Ciechanover, 1998). SUMO on the other hand plays a different role unrelated to protein stability, rather it regulates the function of proteins it is ligated to by a number mechanisms (Miura *et al.*, 2007). The dominant role for SUMO is to recruit interactions with other proteins by binding to short SUMO interacting motifs (SIMs) within other proteins (Hecker *et al.*, 2006). SUMO can also regulate protein function by blocking interactions through steric hindrance (Boyer-Guittaut *et al.*, 2005) or through conformational changes to modified protein (Ulrich, 2005). The modification of proteins by SUMO occurs at a lysine residue and in most cases the lysine residue lies within the conserved motif ΨKx[ED]. Additionally, like ubiquitin, poly-SUMO chains can be formed by the SUMOylation of lysines within SUMO itself.

In *Arabidopsis thaliana* (*Arabidopsis* hereafter) there are eight SUMO paralogs though only *SUMO1*,*2*,*3* and *5* (*SUM1*-3 & 5) are expressed at detectable levels (Budhiraja *et al.*, 2009), suggesting the remaining SUMO paralogs may be pseudogenes or do not play a significant role within the cell. The SUMO paralogs SUM1 and SUM2 are the most similar, they are expressed at the highest levels and they have a high degree of functional redundancy. SUM3 is more divergent and SUM5 is the most dissimilar of the four paralogs. Single mutants of either *SUM1* or *2* do not show a phenotype while the double mutant *sum1 sum2* is lethal (Saracco *et al.*, 2007). SUM1 and 2 appear to play a dominant role in plant physiology and are preferentially conjugated to targets over SUM3 or SUM5 (Castaño-Miquel *et al.*, 2011). The expression profiles for the various SUMO paralogs are different suggesting that in plants the different paralogs have distinct functional roles (van den Burg *et al.*, 2010).

1.1 The SUMOylation cascade

Both SUMO and ubiquitin are attached to proteins by a similar cascade of enzyme reactions, with specific enzymes for both the SUMO and ubiquitin cascades (Miura & Hasegawa, 2010). SUMO proteins are expressed as pro-peptides that undergo post-translational processing to produce the mature species

by SUMO proteases which cleave a small fragment off the C-terminal end of the protein. Mature SUMO terminates with a di-glycine motif at its C-terminal end and ligation to SUMO targets occurs at the terminal glycine of this motif forming an isopeptide bond with a lysine side chain in the targets. The di-glycine motif in SUMO is essential for ligation to target lysine residues. Ligation occurs through a cascade of reactions mediated by three classes of enzyme namely E1, E2 and E3 shown in Figure 1.1.

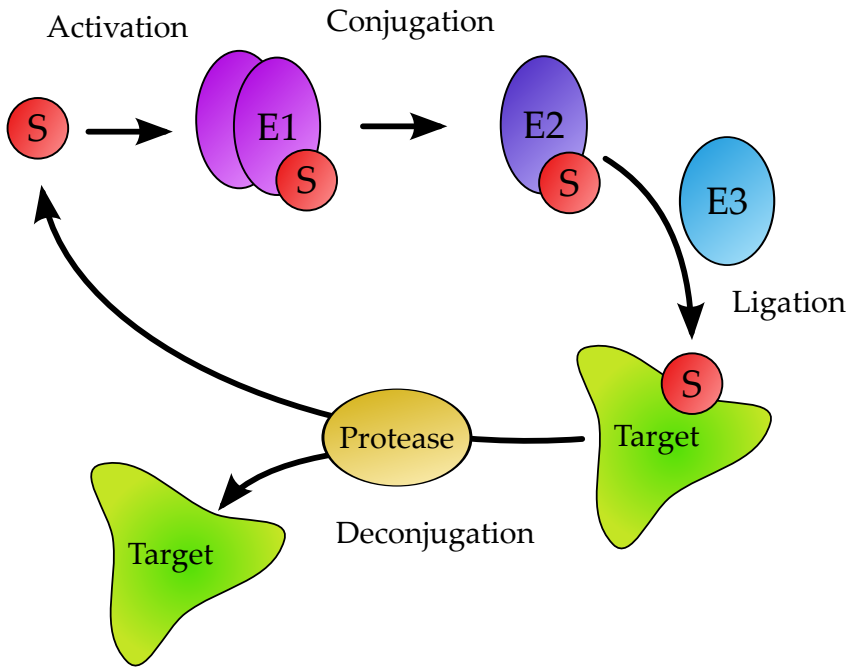


Figure 1.1: The SUMOylation enzyme cascade. SUMO (S) is activated by an E1 activating heterodimer then transferred to an E2 conjugating enzyme then ligated to a target. SUMO E3 ligases may assist in SUMOylation however, SUMOylation can still occur in the absence of an E3. SUMO proteases can remove SUMO from conjugated targets and the SUMO is then recycled back into the pathway.

The SUMO activating enzyme (SAE) or E1 catalyses the ATP dependant process of activation whereby the E1 captures free SUMO by forming a covalent thioester bond to the SUMO molecule, generating activated SUMO. The E1 is a heterodimer of two proteins SAE1 and SAE2 and in *Arabidopsis* there are two isoforms of SAE1, SAE1a and SAE1b. The activated SUMO is then passed to a SUMO conjugating enzyme (SCE) or E2 by transfer of the thioester linkage to the E2. In the final step E3 ligases mediate the conjugation to a target, however, while an E3 is essential for ubiquitin ligation, SUMO ligation can occur in the absence of an E3. SUMO E3 enzymes have been shown to enhance the rate of SUMOylation by binding to and restricting the flexibility of the thioester linkage between the SUMO and E2 enzyme which maintains a more favourable conformation for SUMO ligation (Truong *et al.*, 2011). The mechanism of action for the SUMO E3 ligases is different to that of ubiquitin E3 ligases which bind to specific motifs or domains within the target proteins and thus confer substrate specificity, something that SUMO E3s do not. It is also of interest to note that while there are at least 1400 ubiquitin E3 ligases in *Arabidopsis* (Sadanandom *et al.*, 2012), only four SUMO E3 ligases have

been found to date (Novatchkova *et al.*, 2012). While there are only four published plant SUMO E3 ligases, it is possible that more exist as there are an appreciable number in other eukaryotes (Wilkinson & Henley, 2010). The four plant E3s were discovered through homology searching with animal and yeast E3s. It may be the case that there are classes of plant specific E3s that have little or no homology with animal or yeast and have yet to be identified.

Like ubiquitin, chains of poly-SUMO can form on a target protein. In plants AtSUM1 and AtSUM2 can form chains while SUM3 cannot as it lacks a SUMOylatable lysine residue (Chosed *et al.*, 2006) and it has been proposed that AtSUM3 could act as a SUMO chain terminator (Ulrich, 2008) *in planta*. The purpose of these chains appears to be to enhance the strength of an interaction with SUMO binding proteins (SBPs). SBPs often contain regularly spaced repeats of SIMs in their amino acid sequence, with each SIM binding to a SUMO molecule in a poly-SUMO chain. Poly-SUMO chains binding to tandem repeats of SIMs in this way enhance the strength of the interaction between the two interacting proteins and the number of SUMO units in a chain could be used to regulate interaction strengths. More recently, hybrid poly-SUMO-ubiquitin chains have been found and these have been observed in two distinct instances. The first is formed when SUMO targeted ubiquitin ligases attach a chain of ubiquitin to a growing SUMO chain leading to proteasomal degradation of the target. This mechanism allows SUMOylation to target proteins for degradation. Secondly, hybrid SUMO-ubiquitin chains can act as distinct signals and proteins with hybrid SIM and ubiquitin interacting motif repeats have been identified that bind specifically to these hybrid chains (Guzzo & Matunis, 2013).

After ligation of SUMO, various families of SUMO proteases can later cleave attached SUMO molecules from SUMOylated proteins, returning the SUMO protein to the pool of free SUMO (Miura *et al.*, 2007). These proteases, along with the ligation machinery, allows SUMO to be transiently attached to target proteins, allowing for many levels of regulation. Antagonisation between the SUMOylation cascade and SUMO deconjugation machinery allows for the level of SUMOylated targets in a cell to be regulated and a shift in the levels of these components can lead to a shift in the levels of SUMOylated proteins in the cell. The importance of the de-SUMOylating components has been highlighted by knock-outs of SUMO protease genes which show strong phenotypes (Reeves *et al.*, 2002). Double mutants of the proteases *OVERLY TOLERANT TO SALT1* and *2* (*OTS1* & *2*), *ots1 ots2*, show a dwarf phenotype that is hypersensitive to salt stress and higher levels of SUMOylated protein levels than wild type plants (Conti *et al.*, 2008).

1.2 SUMO interacting motifs

The dominant mechanism by which SUMO modulates protein function is through interaction with SIMs in other proteins. The β-grasp fold structure of SUMO consists of an α-helix and a β-sheet and this

structure is important for forming interactions with SIMs. SIMs tend to be disordered stretches of a peptide lacking secondary structure (Vogt & Hofmann, 2012). The SIM binds to SUMO by inserting into a groove between the β-sheet and α-helix, extending the β-sheet in SUMO in either a parallel or antiparallel manner (Song *et al.*, 2004). The core amino acids of the SIM inserting into this groove and participating in the β-sheet are mostly hydrophobic (typically valine, leucine and isoleucine). This core is flanked by polar or charged amino acids that interact with portions of SUMO outside the groove (Namanja *et al.*, 2012). SIMs often show SUMO isoform specificity, binding to one isoform more strongly than another. The differences in SIM binding between different SUMO isoforms allows the functional specialisation of SUMO paralogs.

SIMs play an important role in progression of SUMO through the SUMOylation cascade, with SIMs found in both E2 and E3 enzymes. The role of the SIM in the E2 is to form a complex with an activated E1-SUMO complex which allows transfer of SUMO to the E2 (Duda *et al.*, 2007). SUMO E3 enzymes which can also contain SIMs, can then bind to the SUMO-E2 complexes to facilitate the SUMOylation of a target protein (Yang & Sharrocks, 2010). SIMs also play a role in recognition of target proteins, SIMs located near to SUMO sites in target proteins can bind to SUMO-E2 complexes promoting SUMOylation of the target (Lin *et al.*, 2006). SIMs in target proteins can determine SUMO paralog selection via SIM specificity for a particular SUMO isoform (Tatham *et al.*, 2005).

1.3 Biological role of SUMO

SUMOylation has been shown to be critical for a vast range of physiological and developmental processes and a large number of proteins have been shown to be targets for SUMOylation. Large numbers of proteins involved in protein stability, metabolism and transcription have been shown to be potential targets, though the function of most SUMO targets remains unknown (Elrouby & Coupland, 2010). Mutants of components of the SUMO conjugation cascade or deconjugation enzymes show severe phenotypes and have been instrumental in elucidating the role of SUMOylation in plants. While double knockouts of *AtSUM1* and *AtSUM2* are embryo lethal, so are knockouts of components of SUMO E1 and E2 enzymes demonstrating that SUMO conjugation is essential for plant viability (Saracco *et al.*, 2007). The SUMO E3s SIZ and MMS21, while not the only E3 ligases in plants, play a predominant role. Single knockouts of either show a strong phenotype which cannot be complemented by overexpression of the other and the double knockout *siz1 mms21* is embryo lethal (Ishida *et al.*, 2012) showing that these two E3s are essential for viability and have divergent roles in plant physiology. *siz1* mutant plants show a dwarf phenotype with reduced fertility and a range of other abnormalities (Ling *et al.*, 2012). SIZ1 has been found to negatively regulate salicylic acid (SA) signalling and the levels of this hormone are elevated in *siz1* plants. The phenotype of *siz1* plants can be mostly reversed by reducing

SA levels by expressing the bacterial SA biosynthesis inhibitor *nahG,* showing that the observed *siz1* phenotype is predominantly due to SA accumulation (Miura *et al.*, 2010). The phenotype seen with *mms21* mutants on the other hand is not due to accumulation of SA but rather appears to be due to cytokinin signalling instead as cytokinin regulated genes show reduced expression in the *mms21* mutant (Huang *et al.*, 2009). *mms21* mutant plants show reduced fertility, reproductive organ deformities and chromosome mis-segregation and fragmentation during meiosis (Liu *et al.*, 2014).

Knockouts of the SUMO proteases also show strong phenotypes and display higher levels of SUMOylated proteins compared to wild type plants. The higher levels of SUMOylated proteins show that SUMOylation is a dynamic process and that transient attachment of SUMO is required for normal cellular function. The SUMO protease mutant *esd4* shows a dwarf phenotype with various developmental abnormalities (Reeves *et al.*, 2002), while double knockouts of the proteases *ots1* and *ots2* show sensitivity to salt stress. Interestingly overexpression of OTS1 in wild type plants increased tolerance to salt stress (Conti *et al.*, 2008).

Cellular plant stresses lead to an increase in SUMOylated protein levels in the cell, indicating that SUMO is implicated in stress response mechanisms. Stresses such as drought, salt, high temperatures and exposure to reactive oxygen species have all been shown to induce SUMOylation in *Arabidopsis* (Yoo *et al.*, 2006; Conti *et al.*, 2008; Kurepa *et al.*, 2003). The E3 MMS21 appears to be a negative regulator of drought tolerance and during drought conditions the expression of *MMS21* decreases. Mutant *mms21* plants show higher drought tolerance than wild type plants while overexpressors show reduced tolerance. MMS21 acts by repressing the expression of stress genes through abscisic acid signalling (Zhang *et al.*, 2013). Conversely the E3 ligase SIZ1 is a positive regulator of drought tolerance and partially responsible for the large increases in SUMOylated protein levels observed during drought stress (Miura & Nozawa, 2014). SIZ1 enhances drought tolerance by positively regulating a number of drought tolerance related genes (Catala *et al.*, 2007) and may regulate drought tolerance through SA mediated stomatal closure which reduces water loss (Miura *et al.*, 2012). Proteomic analyses of SUMOylation during heat shock treatment has revealed that stress induces an increase in SUMOylation of the same set of proteins that are SUMOylated under normal conditions rather than the SUMOylation of new targets. A large proportion of the proteins that showed the highest increase in SUMOylation were RNA and chromatin related genes as well as the heat shock protein transcription factor HSF2A. This suggests that the observed increase in SUMOylation leads to differential gene expression (Miller *et al.*, 2010). The E3 SIZ1 confers tolerance to elevated levels of copper and is required for correct distribution within plant tissues and mutant *siz1* plants display copper toxicity at lower levels than wild type plants (Chen *et al.*, 2011). These data suggest that SUMOylation has a role in heavy metal tolerance however, to date no research exists on the role of SUMOylation in response to metals other than copper.

The E3 ligase SIZ1 has also been implicated in nutrient assimilation processes. Phosphate starvation

leads to changes in root architecture through induction of lateral roots in order to increase phosphate uptake. SIZ1 is a negative regulator of this process and acts by negatively regulating auxin patterning. During phosphate starvation the levels of SIZ1 protein decline and this leads to the expression of various genes associated with root development (Miura *et al.*, 2011) as well as genes required for phosphate uptake, transport and assimilation (Miura *et al.*, 2005). Conversely SIZ1 positively regulates nitrogen assimilation. The nitrate reductase enzymes NIA1 and 2 are SUMOylated, which is mediated by SIZ1, and SUMOylation of these nitrate reductase enzymes enhances their activity. These enzymes are required for nitrate processing and *siz1* plants show an accumulation of nitrate and reduction in other nitrogen containing molecules as well as a nitrogen starvation phenotype compared to wild type plants (Park *et al.*, 2011).

While it has been shown that SIZ1 is vital for correct SA signalling, SUMO also plays a role in abscisic acid signalling (ABA) (Lois *et al.*, 2003).The SUMOylation of the transcription factors MYB30 and ABI5 involved in ABA signalling inhibits their activity. ABI5 is one of the many ABA responsive transcription factors acting as a hormone receptor (Finkelstein & Lynch, 2000) which is stabilised in the presence of ABA. SUMOylation of ABI5 also stabilises the protein leading enhanced ABA signalling (Miura *et al.*, 2009). MYB30 on the other hand is not involved in ABA perception but knockouts show ABA hypersensitivity. SUMOylation of MYB30 is required for normal function and the non-SUMOylatable myb30 K38R mutant shows a partial ABA hypersensitivity phenotype (Zheng *et al.*, 2012). Apart from the hormones SA and ABA, the role SUMOylation plays with the other major plant hormones remains unclear, however, part of the work presented in this thesis in chapters 2.3.12&5 demonstrates that SUMO negatively regulates gibberellic acid signalling during stress.

A range of developmental processes are governed by SUMO and mutants of any of the SUMO cascade enzymes show growth and development effects. Mutant *siz1* plants show a range of developmental abnormalities which have been investigated in detail. The increased levels of SA in this mutant lead to reduced plant size and at least part of the reduction in growth is due to inhibition of cell elongation and division which is governed by a number of SA regulated xyloglycan endotransglycosylase/hydrolase genes (Miura *et al.*, 2010). Floral development is severely affected in *siz1* mutants though not all defects are due to the increased levels of SA in these mutants. Female gametophyte development is impaired leading to reduced seed yield and this effect is not reversed by reducing SA levels in this mutant (Ling *et al.*, 2012). SIZ1 regulates flowering time through the floral repressor protein FLC. The regulation of this protein by SUMO has a rather interesting mechanism with three levels of SUMO regulation. SUMOylation of FLC is required for the normal floral repression function of this protein. SIZ1 is able to bind to FLC and surprisingly this *reduces* the rate of SUMO conjugation to FLC. Also, by binding to FLC, SIZ1 has the effect of stabilising the FLC protein (Son *et al.*, 2014). This complicated mechanism in turn regulates flowering time by modulating FLC activity and stability.

SUMOylation plays a role in defence mechanisms against plant pathogens. SIZ1 negatively regulates innate immunity against biotrophic pathogens through systemic-acquired resistance (SAR) resulting from increased SA levels. Mutant *siz1* plants, which hyperaccumulate SA, show constitutive SAR and increased expression of pathogenesis-related genes and exhibit enhanced resistance against the biotrophic pathogen *Pseudomonas syringae* pv. tomato DC3000. Resistance to the necrotophic pathogens mediated by the jasmonic acid pathway appears to regulated independently of SIZ1 as *siz1* plants show no change in susceptibility to the necrotrophic pathogen *Botrytis cinerea* (Lee *et al.*, 2007). Whether SUMOylation plays a role in immunity against necrotrophic pathogens is uncertain. The pathogen *Xamthamonas campestris* (*Xc*) has evolved a SUMO protease effector protein, XopD, that is injected into host cells through the bacterial type III secretion system and weakens the host defence against the pathogen. XopD leads to reduced SA levels and differential gene expression however, the mechanism of XopD function remains unclear but it is possible that XopD de-SUMOylates some as yet unidentified protein in the plant cell to elicit its effect (Kim *et al.*, 2008).

DNA damage is a common occurrence resulting from exposure to various mutagens including reactive chemical species, ultraviolet radiation, ionising radiation and replication errors. Cellular homeostasis relies on effective DNA repair mechanisms to correct these errors which would otherwise lead to the loss genome integrity. A multitude of repair processes have developed that repair the various kinds of DNA damage. Various types of damage can occur and different mechanisms of repair are used to address them. DNA double strand breaks (DSBs) are the most severe type of damage and these are repaired either through nonhomologous end joining or homologous recombination. Nucleotides can be modified by various mutagens leading to the formation of DNA adducts which are repaired by excision of the faulty nucleotide leading to a single strand break which is then repaired. Nucleotide excision is also used to repair nucleobase mismatches. Overall the process of DNA repair involves a large number of enzymes and regulatory proteins and the posttranslational modification of these components by ubiquitination, phosphorylation and SUMOylation is used to regulate their activity (Jackson & Durocher, 2013). The importance of SUMO in DNA repair in mammalian systems was noted with the discovery that HsSUM1,2 and 3 as well as SUMO E2 and E3 enzymes accumulated around DSBs (Galanty *et al.*, 2009; Morris *et al.*, 2009). SUMO plays an important role in forming the protein repair complexes required for DNA repair. A mutation in one repair component in mammalian systems, BLM, that leads to the loss of its SUMO site results in increased sensitivity to DNA damage (Ouyang *et al.*, 2009). SUMO regulated DNA mechanisms are also present in plants and are most important for maintaining the integrity of the meristem cell pool which develops into new plant organs during growth. The E3 ligase MMS21 is important for the repair of DNA DSBs and mutants of this protein lead to sensitivity to DNA damaging agents. The DNA repair targets of MMS21 remain unknown but the protein has been shown to associate with the chromatin maintenance proteins MAINTENANCE OF CHROMOSOMES5 and 6

(Xu *et al.*, 2013). A number of chromatin regulation proteins have been identified that have functional SIMs that include DNA and histone methyltransferases and demethylases suggesting that SUMO plays a role in *Arabidopsis* DNA methylation processes, however, the direct role and mechanisms of these proteins have not yet been elucidated (Elrouby *et al.*, 2013). DNA repair mechanisms in plants are less well understood than in animal systems and the role of SUMO even less so, however, recent research has shown SUMO is likely to play an important role as it does in animal systems.

1.4 Purpose of work

The purpose of the work presented in this thesis was to investigate the role of SUMOylation in the model plant *Arabidopsis* with a focus on the mechanisms and role of the SUMO-SIM interaction in the plant AtSUMO1 (AtSUM1) protein. The peptide sequence requirements for the interaction were investigated and compared a with a human homolog, HsSUMO1 (HsSUM1), to investigate to what degree the binding properties of SUMO proteins from different species vary. A peptide library was used to screen a large number of small SIM-like peptides for interaction with AtSUM1 and HsSUM1 and a large dataset of interactions was generated. This dataset was used to build a predictor of AtSUM1 and HsSUM1 SIMs from the primary sequence of proteins using the random forest machine learning algorithm. This work presents the first SUMO isoform specific SIM predictor developed to date. The random forest algorithm has been used in the past to predict SUMO sites in proteins and an improved method to build SUMO site predictors is presented which outperforms previous predictors.

The role of SUMOylation in the gibberellin hormone pathway will also be presented. Recently it was discovered that DELLA repressor proteins in the gibberellin pathway are SUMOylated (Conti *et al.*, 2014) and a model was proposed whereby SUMOylated DELLA proteins negatively regulate gibberellin signalling by binding to the gibberellin receptor GID1. Various aspects of this model were tested and it was shown that the receptor GID1 can bind to SUMOylated proteins, and a potential SIM in the protein was identified demonstrating that the proposed model is plausible.

Chapter 2

Materials and methods

2.1 Materials and reagents

Basic laboratory chemicals for preparing buffers and for protein expression media were purchased from Fisher Scientific. Antibiotics and media for plant tissue culture were purchased from Melford. Reagents purchased from other companies will be noted throughout this chapter.

2.1.1 Antibiotics

Antibiotics were used to select transformed organisms or organisms with selection markers. Antibiotic stock solutions were made up according to Table 2.1 and were sterilised by passing through a 0.2 μm cellulose acetate filter. Antibiotic stock solutions were stored at -20°C. Additionally, zeocin aliquots were wrapped in aluminium foil to protect them from light. All media requiring antibiotic supplementation was prepared just before use. Agar plates supplemented with antibiotics were prepared in advance and were stored at 0-5°C and were not kept for longer than 4 weeks. Autoclaved media was allowed to cool to below 50°C before the addition of antibiotics to prevent degradation of heat sensitive antibiotics.

Antibiotic	**Working concentration (μg/ml)**	**Stock concentration (mg/ml)**	**Storage solvent**
Carbenicillin	50	50	water
Chloramphenicol	34	34	100% ethanol
Gentamicin	20	20	water
Kanamycin	50	50	water
Rifampicin	25	12.5	100% methanol
Spectinomycin	50	50	water
Zeocin	100	100	water

Table 2.1: Antibiotics for selection of transgenic organisms.

2.1.2 Antibodies

Antibodies were used in western blotting for detecting specific proteins. Newly purchased antibodies were separated into 5 µl aliquots and stored at -80°C. Western blotting probing solutions containing primary antibodies were recycled and the probing solutions were stored at -20°C. Solutions were reused up to 5 times.

Epitope	**Type**	**Animal raised in**	**Company/Organisation**	**Product code**
GST	Polyclonal	Rabbit	Sigma Aldrich	G7781
6x His	Monoclonal	Mouse	GE Life Sciences	27-4710-01
HA	Monoclonal	Rat	Roche	11867423001
AtSUM1	Polyclonal	Rabbit	Abcam	ab5316
HsSUM1	Polyclonal	Rabbit	Enzo Life Sciences	BML-PW9460
RGA	Polyclonal	Sheep	Nottingham University	

Table 2.2: Primary antibodies.

Epitope	**Type**	**Animal raised in**	**Company/Organisation**	**Product code**
Rabbit IgG	Polyclonal	Goat	Sigma Aldrich	A0545
Mouse IgG	Polyclonal	Rabbit	Sigma Aldrich	A9044
Rat IgG	Polyclonal	Rabbit	Sigma Aldrich	A5795
Sheep IgG	Polyclonal	Goat	Nottingham University	

Table 2.3: Secondary antibodies. All secondary antibodies were conjugated to horse radish peroxidase enzyme for chemiluminescent detection of the antibodies.

2.1.3 Enzymes, proteins and specialist reagents

Product	Company	Product code
AvrII	New England BioLabs	R0174S
BugBuster® Mater Mix	Merck Millipore	71456
cOmplete, Mini, EDTA-free tablets	Roche	04693159001
DO Supplement –Leu/–Trp	Clontech	630417
DO Supplement –His/–Leu/–Trp	Clontech	630419
DpnI	New England BioLabs	R0176S
HsSUM1 protein	Enzo Life Sciences	UW0150
HyperLadder® 1 kb DNA ladder	Bioline	BIO-33053
Inorganic pyrophosphatase	Sigma Aldrich	I1891
Minimal SD Base	Clontech	630411
Minimal SD Base agar	Clontech	630412
PageRuler Plus Prestained Protein Ladder	Thermo Scientific	26620
Phusion® Hot Start Flex DNA polymerase	New England BioLabs	M0535S
ReddyMix® Master Mix	Thermo Scientific	AB-0575/DC/LD/A
RNase H	New England BioLabs	M0297S
SfiI	New England BioLabs	R0123S
Silwet L-77	Momentive	-
Surfactant P20	GE Life Sciences	BR100054
T4 DNA ligase	New England BioLabs	M0202S
X-α-Gal	Glycosynth	70039
XbaI	New England BioLabs	R0145S
YPD Medium	Clontech	630409
YPER-Plus	Thermo Scientific	78999

Table 2.4: Purchased enzymes, proteins and specialist reagents.

2.1.4 Commercial molecular biology kits

Product name	Purpose	Company	Product code
Amine Coupling Kit	SPR chip coupling	GE Life Sciences	BR100050
DNeasy Plant Mini Kit	DNA extraction from plant tissue	Qiagen	69104
Gateway® LR Clonase® II Enzyme mix	Plasmid construction	Life Technologies	11791-020
pENTR™/D-TOPO® Cloning Kit	Gene cloning	Life Technologies	K2400-20
Plasmid Midi Kit	Plasmid extraction from *E. coli*	Qiagen	12143
QIAprep Spin Miniprep Kit	Plasmid extraction from *E. coli*	Qiagen	27106
QIAquick Gel Extraction Kit	DNA extraction from agarose gel	Qiagen	28704
Spectrum™ Plant Total RNA Kit	RNA extraction from plant tissue	Sigma Aldrich	STRN50
SuperScript® III Reverse Transcriptase	cDNA synthesis	Life Technologies	GBP 43.70
µMACS GST Isolation Kit	Immunoprecipitation of GST tagged proteins	Miltenyi Biotech	130-091-370

Table 2.5: Commercial molecular biology kits.

2.1.5 Plasmids

A range of plasmids were used to express proteins in bacteria, plants and yeast. These are shown in Table 2.6. Most of the plasmids used the Gateway technology from Life Technologies to insert gene fragments into the plasmids using homologous combination. A gene insert can be moved from an entry pENTR plasmid into any destination vector using LR Clonase II enzyme mix from Life Technologies. This system was used to generate most expression vectors.

Name	**Purpose**	**N-tag**	**C-tag**	**Company/citation**
pACYC Duet	*E. coli* dual protein expression	6x His; S	-	Merck Millipore
pCDF Duet	*E. coli* dual protein expression	6x His; S	-	Merck Millipore
pENTR D/TOPO	cloning	-	-	Life Technologies
pDEST 15†	*E. coli* protein expression	GST	-	Life Technologies
pDEST 17†	*E. coli* protein expression	6x His	-	Life Technologies
pDEST 22†	Yeast 2-hybrid	*GAL4* AD	-	Life Technologies
pDEST 32†	Yeast 2-hybrid	*GAL3* DBD	-	Life Technologies
pET DEST 55†	*E. coli* protein expression	Strep II	6x His	Merck Millipore
pEarleyGate 201†	Plant protein expression (35S)	HA	-	Earley *et al.* (2006)
pGAPZα B	*P. pastoris* protein expression	α-factor	myc:6x His	Life Technologies

Table 2.6: Plasmids used for protein expression. The dagger symbol (†) indicates Gateway compatible destination plasmids.

2.1.6 Bacterial strains

Table 2.7 shows the bacterial strains used for plasmid maintenance and protein expression and plant transformation. Chemically competent stocks were prepared from the original commercial samples. *Escherichia coli* and *Agrobacterium tumefaciens* are abbreviated to *E. coli* and *A. tumefaciens* respectively in this chapter.

Species	Bacterial strain	Use	Company / publication
E. coli	DH5α	Plasmid maintenance	
E. coli	BL21 (DE3)	Protein expression	New England BioLabs
E. coli	CodonPlus RIL (DE3)	Protein expression	Agilent Technologies
A. tumefaciens	GV3101	Plant transformation	Koncz & Schell (1986)

Table 2.7: Bacterial strains.

2.1.7 Yeast strains

Yeast strains were used to perform yeast two-hybrid (Y2H) assays and for protein expression. Details of the strains used are shown in Table 2.8.

Yeast species	Strain	Use	Company
Saccharomyces cerevisiae	AH109	Yeast two-hybrid	Life Technologies
Pichia pastoris	SMD1168	Protein expression	Life Technologies

Table 2.8: Yeast strains.

2.1.8 Plant materials

Arabidopsis thaliana ecotype Columbia-0 (Col-0) was used for all plant work. Plants were grown under long day conditions at 22°C. Seed collect from plants was stored either in paper bags or in 1.5 ml tubes and were stored at room temperature in the dark.

2.1.9 Commonly used buffers

ECL

- Solution A: 200 mM Tris, pH 8.5, 800 μM *p*-coumaric acid, 10 mM 3-aminophthalhydrazide.
- Solution B: 0.4% w/v H_2O_2.
- Solutions A and B are mixed together in a 1:1 ratio just before use (final concentration: 100 mM Tris, pH 8.5, 400 μM *p*-coumaric acid, 5 mM 3-aminophthalhydrazide, 0.2% w/v H_2O_2).

TBST

- 50 mM Tris, 150 mM NaCl, 0.1% Tween 20, pH 7.6.

TE buffer

- 10 mM Tris and 1 mM EDTA in pH 8.0.

4x SDS-PAGE sample buffer

- 200 mM Tris, 8% w/v SDS, 40% v/v glycerol, 4% v/v β-mercaptoethanol, 50 mM EDTA, 0.1% w/v bromophenol blue.

2.1.10 Commonly used culture media

All culture media was sterilised by autoclaving at 121°C for 20 minutes.

½ MS agar

½ MS agar was used to culture *Arabidopsis* plants.

- 2.15 g/l Murashige and Skoog Basal Salt Mixture, 5% w/v sucrose, 0.7% phytoagar, pH 7.4.

YPG media

YPG was used for the culture of *P. pastoris*.

- 0.1% w/v yeast extract, 0.1% w/v peptone and 0.1% w/v glycerol.

2.2 Molecular biology methods

2.2.1 Bacterial culture

Both *E. coli* and *Agrobacterium tumefaciens* bacteria were grown on low salt Luria-Bertani (LB) 1.2% agar plates. The agar plates were supplemented with appropriate antibiotics to select bacterial colonies with required markers. The concentrations of antibiotics used are shown in Table 2.1. Bacteria were streaked onto plates using a plastic loop to isolate monoclonal colonies when reviving bacteria from cryogenic stocks and when receiving clones from other researchers. Plates were sealed using Parafilm® to prevent drying out. *E. coli* were incubated for >12 hours or overnight at 37°C for colonies to grow and were then stored at 5°C for short term storage. *A. tumefaciens* plates were incubated at 28°C for 2-4 days for colonies to develop and were then stored at 0-5°C for short term storage. Bacterial plates stored at 0-5°C were not kept longer than 2 weeks.

2.2.2 Bacterial glycerol stocks

Long-term glycerol stocks were made of all unique bacterial clones. To prepare the glycerol stocks, a single bacterial colony was used to inoculate 10 ml of LB supplemented with appropriate antibiotics and grown overnight with shaking at either 37°C or 28°C for *E. coli* or *A. tumefaciens* respectively. 600 µl of the bacterial culture was then transferred to a sterile 1.5 ml tube and 600 µl of sterile 50% v/v glycerol solution was added and mixed by pipetting. The 1.5 ml glycerol stock was then flash frozen in liquid nitrogen and then moved to a -80°C freezer for long-term storage.

To revive bacterial clones, the glycerol stock was removed from the freezer and transported in liquid nitrogen. Glycerol stocks were opened under sterile conditions and a small piece of the frozen bacterial culture was dislodged using a plastic pipette tip and then the stock was returned to the liquid nitrogen container. The frozen culture on the pipette tip was then streaked onto an agar plate with appropriate selection antibiotics and grown until colonies developed.

2.2.3 Preparing chemically competent *E. coli*

To transform *E. coli* with plasmids, chemically competent cells were prepared. To transform more than one plasmid into *E. coli*, as is required for the reconstituted SUMOylation system; repeated rounds of transformation followed by generating chemically competent cells were performed. For lines harbouring one or more plasmids, appropriate selection antibiotics were included in the growth cultures.

Required reagents

- LB media
- Wash solution: 100 mM MgCl2
- Cryo solution: 85 mM CaCl2, 15% glycerol

All solutions were sterilised by autoclaving.

Method

A single colony from an agar plate was used to inoculate 10ml of LB, which was grown overnight at 37°C with vigorous shaking. At the start of the day of cell preparation, the wash and cryo solutions were placed on ice to cool to around 0°C. The 2 ml of the starter culture was used to inoculate 100 ml of LB in a 500 ml baffled Erlenmeyer flask which was incubated at 37°C with shaking at 180 rpm. The bacteria were grown until the OD_{600} reached 0.5. The bacterial culture was then immediately placed in an ice bath and agitated for 2 minutes to cool the bacterial culture. The culture was then transferred into centrifuge tubes, and placed into a centrifuge rotor pre-cooled to 0°C then centrifuged at 5000 g for

10 minutes to pellet the cells. From this point in the protocol every effort was made to ensure the cells remained close to 0°C at all times. The supernatant from the culture was discarded and the cell pellet was placed on ice. 100 ml of pre-cooled wash solution was added to pellet, which was then re-suspended by agitating with a sterile plastic loop. The cells were then left on ice for 30 minutes. The cells were then centrifuged as before and the supernatant discarded. The pellet was then suspended in 10 ml of pre-cooled cryo solution and left to incubate on ice for 1 hour. 100 µl aliquots of the cells were pipetted into 1.5 ml tubes pre-cooled in a -20°C freezer. The aliquots were immediately frozen in liquid nitrogen and then moved to a -80°C freezer for long-term storage.

2.2.4 Transformation of chemically competent *E. coli*

The following protocol was used to transform all *E. coli* strains.

Required materials

- Chemically competent *E. coli* cells
- Plasmid DNA (10 - 200 ng/µl)
- SOC medium (SOB + 20 mM glucose:)
  - First, prepare SOB (0.5% w/v yeast extract, 2% w/v tryptone, 10 mM NaCl, 2.5 mM KCl, 20 mM $MgSO_4$, pH 7.5) and sterilise by autoclaving.
  - Add 10 ml filer sterilised 2 M glucose solution per litre of SOB to make SOC.
- Agar plates with selection antibiotics for plasmid marker gene.

Method

Aliquots of competent cells in 1.5 ml tubes were first thawed on ice for 10 minutes. Up to 2 µl of plasmid DNA was added to the competent cells. The competent cells were then flicked once to mix the DNA and the tube was shaken to collect the cells at the bottom of the tube, which was then placed on ice and left to incubate for 30 minutes. The competent cells were then heat shocked by placing in a 42°C water bath for 25 seconds then promptly moved back onto ice for 5 minutes. 500 µl of sterile, room temperature SOC medium was then added to the cells which were then moved to a 37°C incubator heated and shaken at 220 rpm for 1 hour. 20 µl of the cell mixture was then pipetted onto one half of an agar plate warmed to room temperature. The remaining cells were centrifuged at 5000 g for 30 seconds to pellet the cells. The supernatant was then removed leaving around 50 µl in the tube. The pelleted cells were then suspended in the remaining supernatant and all of the mixture was pipetted on to the other half of the agar plate. Each drop of bacterial culture was spread over its respective half of the agar

plate with a sterile spreader. Using dilute and concentrated mixtures on the same plate allowed single colonies to develop irrespective of whether the transformation efficiency was high or low. Agar plates were then sealed with Parafilm and incubated overnight at 37°C to allow colonies to grow.

2.2.5 Transformation of chemically competent *A. tumefaciens*

Required materials

- Chemically competent *A. tumefaciens* cells
- Plasmid DNA (100 - 200 ng/µl)
- LB medium
- Agar plates with selection antibiotics for plasmid marker gene.

Method

Aliquots of competent *A. tumefaciens* cells in 1.5 ml tubes were thawed on ice for 30 minutes. 5 µl of plasmid DNA solution was then added to the cells. The competent cells were then flicked once to mix the DNA and the tube was shaken to collect the cells at the bottom of the tube, which was then placed on ice and left to incubate for 30 minutes. The cells were then placed in liquid nitrogen for 5 minutes followed by heat shock at 37°C for 5 minutes and were then placed on ice for 5 minutes. 1 ml of LB medium was then added to the cells, which were then moved to a 28°C incubator heated and shaken at 220 rpm for 2 hours. The cells were then pelleted by centrifugation at 5000 g for 1 minute and the supernatant was removed leaving 100 µl behind. The cells were suspended in the remaining supernatant and all of the cell mixture was pipetted on to agar plates and spread over the agar surface with a sterile spreader. The agar plates were then sealed with Parafilm then incubated at 28°C for 2-4 days for colonies to develop.

2.2.6 DNA gel electrophoresis

Agarose gel was prepared for DNA gel electrophoresis using commercial 1x Tris–Acetate–EDTA buffer (TAE) and molecular biology grade agarose. Agarose gel concentrations of 0.6 - 2.0 % w/v were used depending on the size of the DNA fragment to be resolved. The TAE buffer and agarose were mixed in an Erlenmeyer flask and heated to boiling point using a microwave oven. The agarose gel was then placed on a shaker for 5 minutes to cool slightly. 1 µl of 10 mg/ml ethidium bromide solution was added per 100 ml of agarose gel. The gel was then placed in a casting tray with a well comb and allowed to set for 30 minutes. Electrophoresis tanks with an electrode separation of 20 cm were used and electrophoresis was carried out at 120 V. 1 kb HyperLadder® DNA ladder was used to determine

DNA fragment size. Agarose gels were visualised using an ultraviolet transilluminator with a digital camera. 1 µl of DNA loading buffer (0.5% w/v bromophenol blue; 0.5% w/v xylene cyanol blue; 30% glycerol) was added per 10 µl of DNA sample before loading onto the agarose gel.

2.2.7 Analytical PCR

Step	Temperature (°C)	Time	Number of cycles
Initial denaturation	95	1 min	1
Denaturation	95	10 s	35
Annealing	Variable	20 s	
Extension	72	1 min/kbp	
Final extension	72	5 min	1

Table 2.9: Analytical PCR program. Used with ReddyMix and other Taq DNA polymerase enzymes.

For routine PCR for testing the presence of a DNA fragment in a sample the following protocol was followed. 2x ReddyMix PCR Master Mix from Thermo Scientific was used. ReddyMix contains DNA polymerase, buffer and dNTPs therefore only DNA template and primers needed to be added. 10 µl reactions in 0.2 ml PCR tubes were used. For each reaction, the following were added: 5 µl ReddyMix, 0.5 µl of each DNA primer (10 µM), 0.5 µl DNA template and 3.5 µl water. Control reactions were always included; water was used for the negative control and a known template for the positive control. PCR reactions were then mixed shaking the tubes and the tubes were centrifuged briefly to collect the PCR reaction mix at the bottom of the tubes.

The optimal annealing temperature was determined for each primer pair. The PCR was performed in a heat cycler with a heated lid. The program used in the heat cycler is shown in Table 2.9.

2.2.8 Bacterial colony PCR

Bacterial colony PCR was used to test for the presence of a DNA fragment in both *E. coli* and *A. tumefaciens.* This technique was used to confirm the correct recombination of vectors using the Gateway system during plasmid construction using D_AttB-F and D_AttB-R primers (see Table A.1 in Appendix A for primer sequences).

For each construct, 5 or more colonies were typically tested for the correct insert. For cloning 10 colonies were typically tested. For each colony a 1.5 ml tube was labelled and 20 µl of sterile water was added. Under a Bunsen burner flame, a single bacterial colony was picked using a 0.5 µl plastic loop and mixed with the water in the 1.5 ml tube. For *A. tumefaciens* half of the bacterial suspension was transferred into a new 1.5 ml tube with identical label. The additional set of tubes was then placed

in a boiling water bath for 5 minutes to lyse the cells. This step was not necessary for *E. coli* cells. PCR reactions were set up according to the analytical PCR method using 0.5 µl of bacterial suspension as DNA template. The PCR were products then analysed by agarose gel electrophoresis. If suitable colonies were identified, the original bacterial suspensions in water were used to inoculate a bacterial culture.

2.2.9 High fidelity PCR

High fidelity PCR using a DNA polymerase enzyme with proof reading was used for cloning and site directed mutagenesis. Phusion® Hot Start Flex DNA polymerase from New England BioLabs was used for all high fidelity PCRs. 50 µl reactions in 0.2 ml PCR tubes were used for PCRs. For each reaction the following components were added: 10 µl of 5x Phusion HF buffer, 1 µl dNTPs (10 mM for each dNTP), 2.5 µl of each primer (10 µM), 2.5 µl DNA template and water to 50 µl. For difficult templates, DMSO was added to a final concentration of 3%. For difficult templates with high GC content, the alternative 5x GC Phusion buffer was used. Reactions were mixed using a vortex mixer then centrifuged to collect reaction mix at the bottom of the tube. PCRs were performed on a heat cycler with a heated lid using the program shown in Table 2.10.

Step	**Temperature (°C)**	**Time**	**Number of cycles**
Initial denaturation	98	10 s	1
Denaturation	98	5 s	35
Annealing	Variable	10 s	
Extension	72	15 s/kbp	
Final extension	72	5 min	1

Table 2.10: High fidelity PCR program. This program was used with Phusion® DNA polymerase.

2.2.10 cDNA synthesis

RNA was collected from whole *Arabidopsis* plants to be used as a substrate for cDNA synthesis. Col-0 seeds were surface sterilised then spread onto sterile filter paper soaked with ½ MS media with 5% w/v sucrose in a square petri dish. The seeds were vernalised overnight then moved to an incubator with a long day light cycle. After 9 days of growth, the plants were removed from the surface of the filter paper and then rinsed with distilled water. RNA was extracted from the plants using a Spectrum™ Plant Total RNA Kit as per the kit instructions. DNase treatment of the RNA was not performed since the cDNA was only used for cloning and not gene expression quantification. The RNA was then used as a template for cDNA synthesis using SuperScript® III Reverse Transcriptase as per the included instructions for

the enzyme. 20-mer oligo dT was used to prime the RNA for cDNA synthesis from mRNA. The cDNA was eluted into 20 µl and then 65 µl of distilled water, 10 µl of 10x RNase H buffer and 5 µl of RNase H (25 U) was added to the cDNA to digest the RNA. The cDNA was incubated with RNase H at 37°C for 20 minutes followed by treatment at 65°C for 20 minutes to deactivate the RNase H enzyme. The cDNA was stored at -20°C.

2.2.11 Blunt end cloning into pENTR D/TOPO plasmids

Full length coding DNA sequences (CDS) lacking a stop were cloned into pENTR D/TOPO plasmids. Forward primers were designed that add the 'CACC' DNA sequence onto the 5' end of the CDS and reverse primers were designed that exclude the stop codon at the 3' end of the CDS. Primers used for cloning are shown in Table A.1 in the appendices (primer prefixed with 'C_'). CDS fragments were amplified from cDNA using the high fidelity PCR method. The amplified DNA was then separated on an agarose gel by electrophoresis. The correct size DNA fragment was excised from the agarose gel using a scalpel. The DNA band was illuminated during excision using a UV transilluminator with a perspex UV shield. DNA was then extracted from the agarose gel sample using a Qiagen gel extraction kit according to the manufacturer's instructions, including an optional wash step to remove excess salts. The concentration of the recovered DNA sample was measured using a NanoDrop spectrophotometer.

The purified DNA fragment was then cloned into the pENTR D/TOPO plasmid using the pENTR D/TOPO cloning kit as per the instruction manual. The resulting product from the kit was transformed into Top10 *E. coli* cells. 10 colonies for each cloned gene were screened using colony PCR to check for colonies with the correct insert size. If suitable colonies were identified, they were used to inoculate 10 ml LB cultures which were grown overnight at 37°C with shaking. Plasmid DNA was extracted from these cultures using a QIAprep Spin Miniprep Kit according to the manufacturer's instructions. Isolated plasmid DNA was then sequenced using the in house sequencing service at Durham University. M13 (-20) forward and M13 reverse primers were used to sequence the insert. If clones without errors were not identified, more clones were screened until an error free insert was isolated.

2.2.12 Site directed mutagenesis

Site directed mutagenesis of circular plasmids was performed using PCR adapted from the method by Weiner *et al.* (1994). A forward mutagenic primer was designed to include the required base pair changes then the reverse complementary primer was created. Plasmid DNA of the target was diluted to 2 ng/µl and PCR was set up according to the high fidelity PCR method except that the number of cycles was reduced to 15. 1 µl of the restriction enzyme DpnI was then added directly to the PCR reaction, which was mixed and moved to a clean 0.2ml PCR tube. The reaction was incubated at 37°C for 1 hour and transferred to a new tube and incubated once more at 37°C for 1 hour. The enzyme was

then deactivated by heating the reaction to 80°C for 20 minutes. 2 µl of the product was then used to transform DH5α cells. Plasmid was prepared from the resulting colonies and sequenced to confirm the presence of the mutation.

2.2.13 Restriction digest cloning into pGAPZα B plasmids

The DNA fragments *RGA*, *rga K65R* and *S1:rga K65R* were cloned into the *P. pastoris* expression vector pGAPZα B. The gene fragments were amplified from pENTR D/TOPO plasmids using the high fidelity PCR method. Forward primers introduced a SfiI restriction site onto the front of each gene. The reverse primers for the fragments lacked a stop codon and introduced an XbaI site at the end of each gene. For the *RGA* and r*ga K65R* fragments the primers E_RGA-F and E_RGA-R were used. For *S1:rga K65R* primers E_SUM1-F and E_RGA-R were used (see Table A.1 in the appendices for primer sequences). The gene fragment and pGAPZα B were then double digested with SfiI and XbaI. 42 µl of the PCR reactions and 20 µg of pGAPZα B in 42 µl of water had the following added to them: 1 µl SfiI, 2 µl XbaI and 5 µl 10x CutSmart buffer (New England BioLabs). The reaction was incubated at 37°C for 4 hours followed by 50°C for 1 hour to inactivate the restriction enzymes. The restriction digests were then gel purified using Qiagen gel extraction kit according to the manufacturer's instructions. The purified DNA was quantified using a NanoDrop spectrophotometer. Ligation reactions were set up using 150 ng of digested pGAPZα with each of the digested gene fragments. 247 ng of *RGA* and *rga K65R* and 290 ng o*f S1:rga K65R* was used. DNA plus 2 µl of 10x T4 DNA ligase buffer, 1 µl T4 DNA ligase and water to 20 µl was mixed together for each ligation reaction. The ligation reactions were incubated at 25°C for 2 hours. 1 µl of each ligation reaction was then used to transform competent DH5α cells and colonies were allowed to develop on 50 µg/ml zeocin agar plates which were incubated in the dark at 37°C overnight. Colony PCR was performed on 10 colonies for each construct to select for the correct size gene. The original primers used to amplify the DNA fragments from pENTR D/TOPO were used in the colony PCR. 5 colonies with the correct size insert were then used to set up 10 ml LB cultures supplemented with 50 µg/ml zeocin. The cultures were grown overnight at 37°C with shaking and then plasmid DNA was purified. An aliquot of each sample was tested with SfiI and XbaI to check the size of the insert and the plasmid. Plasmid DNA was then sequenced and clones with the correct DNA sequence were isolated.

2.2.14 Transformation of *P. pastoris*

Plasmid DNA for the transformation was prepared using a Qiagen Midi prep kit. 10 µg of plasmid DNA was then digested with 5 µl of AvrII (25 U) in a 250 µl reaction to linearise the plasmid DNA. The reactions were incubated at 37°C for 4 hours and then the DNA was purified using ethanol precipitation. To each sample 25 µl of 3M sodium acetate and 750 µl of 100% ethanol was added. The samples were

mixed and then left overnight in freezer at -20°C to precipitate the DNA. The plasmid DNA was then pelleted in a refrigerated centrifuge at 0°C for 30 minutes at 14000 g. The supernatant was discarded and 250 μl 70% ethanol was added to wash the DNA. The DNA was pelleted by centrifuging at 14000 g for 5 minutes and the supernatant discarded. The tubes were left open at room temperature for the remaining ethanol to evaporate and the DNA was resuspended in 10 μl of TE buffer with a final concentration of around 2 μg/μl.

Electrically competent *P. pastoris* cells were then prepared. A 5 ml *P. pastoris* starter culture in YPD was grown overnight in a 50 ml Erlenmeyer flask at 30°C. This was used to inoculate a 500 ml YPG culture in a 2 l baffled Erlenmeyer flask, which was grown at 30°C with shaking at 160 rpm. When an OD_{600} of 1.5 was reached, the cells were chilled in an ice bath for 5 minutes then pelleted by centrifugation in a chilled centrifuge at 1500 g for 5 minutes. The pellet was re-suspended in 500 ml of ice-cold water. The cells were pelleted as before and then suspended in 20 ml of ice-cold 1 M sorbitol solution. The cells were pelleted one final time and then suspended in 2 ml of ice-cold sorbitol and then kept on ice until they were used for transformation.

For transformation 80 μl aliquots of cells were transferred to chilled electroporation cuvettes with a gap of 0.2 cm. 10 μg of linearised plasmid DNA was then added to each cuvette and mixed by pipetting. Cells were then placed in a BioRad MicroPulser™ Electroporator and pulsed with the default settings for *S. cerevisiae* (1 pulse of 3 kV). 1 ml of ice-cold 1 M sorbitol was then added to the cuvette and the cells were then incubated at 30°C for 2 hours. The cells were then spread on YPG agar plates with 100 μg/ml of zeocin and left to incubate at 28°C for 3 days in the dark for colonies to develop. Once colonies had formed, 5 for each construct were selected and re-streaked onto new agar plates to isolate monoclonal colonies.

2.3 Protein methods

2.3.1 *E. coli* protein expression

E. coli were cultured in LB for protein expression. An overnight starter culture inoculated from a single colony was grown overnight at 37°C with shaking with appropriate antibiotic expression. The main expression culture was inoculated with 1/20 media volume of starter culture. Cultures were generally grown in baffled Erlenmeyer flasks with 5 times the volume of the expression media, except for 500 ml expression media, which was grown in 2 l flasks. The bacterial cells were grown at 37°C with shaking until the OD_{600} reached 0.6 to 1.0. Media was then cooled on ice to the correct expression temperature which varied depending on the protein being expressed. Media was supplemented with 0.1 mM isopropyl β-D-1-thiogalactopyranoside (IPTG) to initiate protein expression. Depending on temperature, the cells were cultured from 1 to 12 hours. After expression, the cells were chilled in

an ice bath then pelleted in a refrigerated centrifuge at 5000 g for 10 minutes. The supernatant was discarded and the pellet was retained for further processing. At times the pellets were stored for later processing, in this case they were frozen in liquid nitrogen then stored at -80°C until they were needed.

To prepare bacterial pellets for purification, the pellet was weighed and for each gram of bacterial pellet, 5 ml of BugBuster Master Mix supplemented with 1x Complete EDTA free protease inhibitor tablets (Roche) was used. The pellet was suspended in the BugBuster and then the cells were chemically lysed by gently agitating for 15 minutes at room temperature. The insoluble debris was removed from the lysate by centrifugation at 27 000 g at 0°C for 15 minutes followed by filtration through a 0.2 μm pore size cellulose acetate filter.

2.3.2 *Pichia pastoris* protein expression

P. pastoris protein expression was performed using YPG media. Starter cultures inoculated from a single colony were grown for 18 hours at 30°C with shaking. Expression cultures in Erlenmeyer flasks were then inoculated with all of the starter cultures. Small-scale expression used 10 ml of YPG media in 50 ml flasks. Large-scale expression used 500 ml of YPG media in 2 l flasks. Main expression cultures were grown at 30°C for 1-2 days until a high cell density developed. After expression, the cells were chilled in an ice bath then pelleted in a refrigerated centrifuge at 5000 g for 10 minutes. Both the supernatant and cell pellet were retained for protein purification.

To prepare yeast pellets for purification, the pellet was weighed and for each gram of bacterial pellet, 5 ml of Y-PER Plus supplemented with 1x Complete EDTA free protease inhibitor tablets (Roche) was used. The pellet was suspended in the Y-PER Plus and then the cells were chemically lysed by incubating at 45 °C for 10 minutes. The insoluble debris was removed from the lysate by centrifugation at 27 000 g at 0°C for 15 minutes followed by filtration through a 0.2 μm pore size cellulose acetate filter.

2.3.3 Separation of recombinant protein fractions

To analyse recombinant protein solubility in *E. coli* protein expressions, inclusion bodies were separated from the soluble fraction of bacterial lysates. Aliquots of expression media were collected in 1.5 ml tubes. The volume collected was normalised to the number of cells in 2 ml of media at OD_{600} = 1.0. The cells were then pelleted by centrifugation at 5000 g for 5 minutes and the supernatant discarded. Lysis buffer consisting of a solution of BugBuster® Master Mix supplemented with 1x cOmplete, Mini, EDTA-free protease inhibitors was prepared. Bacterial pellets were suspended in 60 μl of lysis buffer and agitated at room temperature for 5 minutes to lyse the cells. The samples were then centrifuged at 13000 g for 10 minutes to pellet the insoluble cell debris. The supernatant was collected and transferred to separate tubes containing 20 μl of 4x SDS-PAGE buffer and mixed. The insoluble debris pellet was

washed by adding 100 µl of inclusion body wash buffer then centrifuged for 1 minute at 13000 g. The wash buffer supernatant was carefully removed leaving the insoluble debris pellet in the tubes. 80 µl of 1x SDS-PAGE sample buffer was then added to the pellet and tubes were tapped against the lab bench to dislodge the pellet. All samples were then placed on heat block at 95°C for 5 minutes, agitating the tubes every minute or so. The samples were then centrifuged at 13000 g for 5 minutes and then either loaded onto an SDS-PAGE gel or frozen at -20°C for later use.

2.3.4 SDS-PAGE

Required solutions

- 1x SDS-PAGE running buffer (24.8 mM Tris, 192 mM glycine, 0.1% w/v SDS)

Method

	Component volume (ml)				
	Resolving gel				**Stacking gel**
Component	**8%**	**10%**	**12%**	**15%**	**5%**
Water	4.6	4.0	3.3	2.3	6.8
30% acrylamide/bis-acrylamide mix	2.7	3.3	4.0	5.0	1.7
1.5 M Tris pH 8.8	2.5	2.5	2.5	2.5	0
1.0 M Tris pH 6.8	0	0	0	0	1.25
10% w/v SDS	0.1	0.1	0.1	0.1	0.1
10% w/v ammonium persulphate	0.1	0.1	0.1	0.1	0.1
TEMED	0.006	0.004	0.004	0.004	0.01
0.5% w/v bromophenol blue	0	0	0	0	0.01

Table 2.11: SDS-PAGE gel components for 10 ml casting solutions.

Sodium dodecyl sulphate polyacrylamide gel electrophoresis (SDS-PAGE) was used to separate and resolve protein samples. The BioRad mini-gel system for gel casting and electrophoresis was used. Casting solutions for resolving and stacking gels were prepared according to Table 2.11 depending on the acrylamide percentage of the resolving gel. Gel casting solutions were prepared on ice to slow polymerisation before they were added to the gel casting plates. Casting plates with a gel thickness of 1 mm were used. First 4.8 ml of resolving gel solution was added to each set of casting plates, then the stacking gel was immediately overlaid onto resolving solution, taking care not to mix the two solutions. A casting comb was then placed in the gel casting plates and the gels were allowed at least 15 minutes to set. Once set, the casting combs were removed and the wells were rinsed with distilled water. The

gels were then placed in an electrophoresis tank and SDS-PAGE running buffer added. Protein samples in SDS-PAGE sample buffer were heated on a heat-block at 95°C immediately prior to loading onto the gels. PageRuler Plus protein ladder was used as a molecular weight marker. Depending on the gel percentage, electrophoresis was carried out between 80 and 140 V until the dye front reached the bottom of the gel. Since the small protein SUMO migrated just behind the dye-front, the dye-front was not allowed to run off the gels to ensure that SUMO protein was not lost.

2.3.5 Western blotting

Required solutions

- 1x transfer buffer (24.8 mM Tris, 192 mM glycine, 10% v/v methanol)
- TBST
- ECL solution

Method

Proteins from acrylamide gels were transferred to polyvinylidene fluoride (PVDF) membranes for western blotting. A PVDF membrane was first soaked in methanol for 5 minutes to wet the membrane. The membrane was then transferred to transfer buffer for at least 5 minutes. An acrylamide gel was placed on the PVDF membrane and sandwiched between filter paper and sponges pre-soaked in transfer buffer. The sandwich was then placed in a transfer cassette and placed in a transfer tank with an ice pack. The tank was filled with ice-cold transfer buffer. Electrophoresis was carried out overnight at 20 V in a cold-room to transfer the proteins. The membrane was then removed and washed in TBST.

The membrane was blocked using TBST with 5% w/v Marvel milk powder for 1 hour on a rocking platform. Primary and secondary antibody solutions were prepared in TBST or TBST milk depending on the antibody. Primary antibody was added directly after blocking and incubated for 1-2 hours. The membrane was then washed 3 times with TBST for 5 minutes. Secondary antibody was incubated for 1 hour and then the membrane was washed again in TBST. The membrane was then placed in 2 ml of ECL solution for 10 seconds then sandwiched between 2 acetate sheets and then exposed to X-ray film. X-ray film was developed in an automated film processor. For protein quantification, the membrane was washed then stained in Coomassie stain to show the immobilised proteins in the membrane.

2.3.6 Coomassie staining

Required solutions

- Coomassie stain (0.1% w/v Coomassie Brilliant Blue R-250, 10% v/v acetic acid, 50% v/v methanol)
- Coomassie destain (10% v/v acetic acid, 40% v/v methanol)

Method

An acrylamide gel or PVDF membrane was placed in a petri dish and covered in Coomassie stain. The gel/membrane was incubated for 30 minutes on a shaking platform then stain removed. The gel/membrane was then rinsed in water until all stain was removed then the Coomassie destain was added to the petri dish to destain overnight. Destain for PVDF membranes was changed once or twice and typically took longer to destain than acrylamide gels.

2.3.7 Protein purification

Column buffers

All reagents were degassed and vacuum filtered through a 0.2 μm pore size cellulose acetate filter.

- His binding buffer (16.2 mM Na_2HPO_4, 3.8 mM NaH_2PO_4, 500 mM NaCl, 30 mM imidazole, pH 7.4)
- His elution buffer (16.2 mM Na_2HPO_4, 3.8 mM NaH_2PO_4, 500 mM NaCl, 500 mM imidazole, pH 7.4)
- GST binding buffer (15.4 mM Na_2HPO_4, 4.6 mM NaH_2PO_4, 140 mM, 2.7 mM KCl, pH 7.3)
- GST elution buffer* (15.4 mM Na_2HPO_4, 4.6 mM NaH_2PO_4, 140 mM, 2.7 mM KCl, 10 mM reduced glutathione, pH 7.3)
- Storage buffer (20% v/v ethanol)

*The GST elution buffer differs from the GSTrap protocol as the buffer was incompatible with downstream SPR experiments since it contained primary amines. The alternative buffer did not contain primary amines.

Affinity column purification

1 ml HisTrap™ HP and GSTrap™ HP sepharose columns were used for purification of recombinant His and GST tagged proteins respectively. The columns were used with either a peristaltic pump (P1,

GE Healthcare) or the ÄKTA FPLC system (GE Healthcare). All actions were performed at 1 ml/min unless otherwise stated. Columns were prepared by pumping 5 ml of distilled water followed by 5 ml of binding buffer. The NaCl and imidazole concentration of the crude protein samples were adjusted to the same concentration as the respective binding buffer. The crude protein samples were filtered through 0.34 μm cellulose acetate filters prior to loading onto the sepharose columns. Crude protein samples were then pumped through the columns. For GSTrap columns, the crude protein samples were pumped at 0.6 ml/min for room temperature purifications; for purifications performed at close to 0°C a flow rate of 0.4 ml/ml was used. The lower flow rates for the GSTrap columns were to account for the slow binding kinetics of the GST-glutathione interaction. Columns were then washed with 10-15 ml of binding buffer. To elute the protein, 15 ml of elution buffer was passed through the columns and 1 ml fractions were collected. For HisTrap columns, a linear gradient elution was used instead. To clean columns, they were washed with 10 ml of water then 5 ml of storage buffer.

The eluted protein fractions were analysed for protein by taking 20 μl aliquots and mixing these with 80 μl Bradford reagent in a white 96 well plate. The colour of the samples was compared to an elution buffer blank. Fractions with protein were pooled and then concentrated and buffer exchanged using Amicon centrifugal columns (Merck Millipore). Protein sample concentrations were measured by infrared spectroscopy using a DirectDetect spectrometer (Merck Millipore). Protein samples were then separated into aliquots and frozen in liquid nitrogen then stored at -80°C. Aliquots were only used once and re-frozen.

Batch affinity resin purification

Batch resin purification was performed using glutathione sepharose 4b resin and HisTrap nickel-nitrilotriacetic acid (Ni-NTA) sepharose resin (GE Life Sciences). The bed volume of the sepharose beads differed between experiments and between 5 and 20 μl was used. All centrifuge steps were performed at 1000 g for 30 seconds. Twice the bed volume of sepharose beads was pipetted into 1.5 ml tubes. The sepharose beads were then washed 5 times with 100 μl of binding buffer, centrifuging between each wash. The protein samples were then applied and the tubes were placed in a rotary mixer for 30 minutes. The sepharose beads were then washed 3-5 times with binding buffer. Samples were then mixed $^1/_3$ bed volume of 4x SDS-PAGE sample buffer and then placed on a heat block heated to 95°C for 5 minutes and then the samples were analysed by SDS-PAGE.

2.3.8 Co-immunoprecipitation of GID1a and SUMO1

Required buffers

- 2x reaction buffer (100 mM Tris, 100 mM NaCl, 0.1% v/v IGEPAL CA-630, 2x Gamborgs B5 basal medium, pH 7.5).

	Component Volume (µl)					
Tube ID	**2x Reaction buffer**	**GST:GID1a**	**GST**	**His:AtSUM1**	**10 mM GA_3**	**Water**
1	500	390	0	100	10	0
2	500	390	0	100	0	10
3	500	0	390	100	10	0
4	500	0	390	100	0	10

Table 2.12: Co-IP of GID1a and SUM1.

Method

Protein samples of GST, GST:GID1a and His:AtSUM1 with a concentration of 1 mg/ml were used. The proteins were in a 15 mM Tris pH 8.0 buffer. The co-immunoprecipitation (co-IP) was performed using paramagnetic GST antibody labelled beads to precipitate free GST and GST tagged proteins. Reactions according to Table 2.12 were set up. The reactions were incubated in ice for 30 minutes for interactions to occur. Miltenyi µ magnetic columns were used for the co-IP. 1x reaction buffer was used as a wash buffer. A wash buffer containing 0.1 mM GA_3 was used for all wash steps with the GA_3 containing reaction (tubes 1 and 3). The columns were placed within a magnetic holder and were equilibrated with 500 µl of the 1x reaction wash buffers. The protein samples were then applied to the columns. Once all the sample had passed through, the columns were washed 4 times with 200 µl 1x reaction wash buffer. A final wash with 100 µl 20 mM Tris pH 7.5 (again with 0.1 mM GA_3 for the respective samples) was performed. A sample of SDS-PAGE sample buffer was then placed on a heat block at 95°C. 20 µl of hot sample buffer was added to each column. The columns were then incubated for 5 minutes. Collection tubes were placed under each column and the protein samples were eluted by the addition of 50 µl of sample buffer. The elutes were then split into two equal aliquots and run two separate SDS-PAGE gels. These gels were then analysed by western blotting with αGST and αAtSUM1 antibodies. Control samples of the input proteins were included.

2.3.9 Yeast two-hybrid assay of GID1a and RGA

Required materials

- 50% polyethylene glycol (PEF) 3350
- 10x TE (100 mM Tris, 10 mM EDTA, pH 7.6)
- 1 M lithium acetate pH 7.6

- 0.8% w/v NaCl
- 1 M 3-amino-1,2,4-triazole (3-AT)
- 20 mg/ml X-α-Gal in DMF
- YPD media
- SD minimal media
- -L/-W amino acid drop out powder
- -L/-W/-H amino acid drop out powder

Method

The genes for GID1a, gid1a V22A and gid1a V22S in pENTR D/TOPO were transferred to pDEST32 using Gateway cloning to make GAL4 DNA binding domain (BD) fusions. The gene for RGA in pENTR D/TOPO was transferred to pDEST22 using Gateway cloning to make a GAL4 activation domain (AD) fusion. The plasmids were used to create 7 different yeast clones each with two plasmids as per Table 2.13. Large amounts of plasmid DNA were purified using a Qiagen Midi-prep kit.

Clone ID	pDEST32	pDEST22
1	RGA	empty
2	empty	GID1a
3	empty	gid1a V22A
4	empty	gid1a V22S
5	RGA	GID1a
6	RGA	gid1a V22A
7	RGA	gid1a V22S

Table 2.13: GID1a-RGA yeast two-hybrid clones.

4 μg of each pDEST22 and pDEST32 was added to labelled 1.5 ml tubes. Transformation solution was made by mixing the following in a separate tube: 2.4 ml 50% PEG, 300 μl 10x TE and 300 1 M Lithium acetate. A yeast suspension was made by mixing 300 μl of water with approximately the same volume of AH109 yeast cells from an agar plate. To each plasmid containing tube 270 μl of the transformation solution and 35 μl of the yeast suspension was added. A vortex mixer was used to homogenise the samples. The cells were then heat shocked at 42°C for 15 minutes then pelleted by centrifuging at 1000 g for 5 minutes. The supernatant was then discarded and the cells were suspended in 500 μl of YPD and left at room temperature for 5 minutes. The cells were pelleted again as before and then suspended in 500 μl of 0.8% NaCl and left overnight in the dark.

The following day the cells were pelleted by centrifuging and 400 µl of the supernatant was removed. The cells were suspended in the remaining 100 µl of supernatant then all of the suspension was spread on double selection -L/-W SD agar plates. The surface of the agar was allowed to air dry then the plates were closed and sealed with Parafilm and then incubated at 28°C for 3-5 days for colonies to develop.

Once colonies had formed, a single colony for each gene construct was mixed with 500 µl of 0.8% NaCl. For the interaction assay 5 µl of these cultures was spotted onto various plates. Double selection -L/-W SD agar plates were used to show that both plasmids were present. Triple selection -L/-W/-H SD agar plates with 40 µg/ml X-α-Gal were used to show interaction and plates with and without 100 µl GA_3 and 5 mM 3-AT were used. This assay was repeated twice more with separate yeast clones.

2.3.10 Surface plasmon resonance of GID1a and SUMO1

Required buffers

- 10 acetate/acetic acid buffers at a range of pH values
- SPR binding buffer (16.2 mM Na_2HPO_4, 3.8 mM NaH_2PO_4, 150 mM NaCl, 0.005% P20, pH 7.4)

Method

Surface plasmon resonance (SPR) was performed using a CM5 amine coupling chip on a Biacore 2000 machine (GE Life Sciences). All reactions took place at 25°C. GST:GID1a, GST:gid1a V22A, GST:gid1a V22S and GST were bound to the CM5 chip and RGA:His and His:AtSUM1 were used as binding ligands. Scouting was performed to find the optimal pH for protein coupling. 10 mM sodium acetate/acetic acid buffers were used for the pH scouting ranging from 4.1 to 5.2 in 0.1 pH steps. 2 µl of each protein sample was mixed with 58 µl of each pH solution and samples were injected at a rate of 20 µl/minute for 1 minute. Once optimal pH values had been determined, each protein was immobilised to the chip surface using an amine coupling kit. 10 µl of protein was used for the immobilisation reaction. The target RU for the GST:GID1a proteins was 2000 response units (RU) and for the GST 800 RU so that there was the same molar ratio between the different proteins.

	Component (µl)			
Reaction ID	**SPR buffer**	**RGA:His**	**His:AtSUM**	**1 mM GA_3**
1	166	0	4	0
2	164.3	0	4	1.7
3	160	10	0	0
4	158.3	10	0	1.7
5	156	10	4	0
6	154.3	10	4	1.7

Table 2.14: GID1a SPR reactions.

Once the GID1a proteins had been immobilised ligation reactions were performed. Reactions were set up according to Table 2.14. Reactions were injected at a rate of 20 µl/min for 2 minutes followed by buffer only at a rate of 20 µl/min for 2 minutes. Each reaction was injected twice. Data was exported and analysed in R. The GST channel was scaled and subtracted from the GID1a channels to subtract non-specific binding to GST protein and the CM5 chip surface.

2.3.11 Reconstituted *E. coli* SUMOylation assay

The reconstituted *E. coli* SUMOylation assay by Okada *et al.* (2009) was used and the plasmids for this assay were obtained from this group. A pACYCDuet plasmid containing *S:SAE1a* and *His:SAE2* was used. A second pCDFDuet plasmid contained one of two *His:AtSUM1* forms and *S:SCE*. One form of SUMO was functional and was denoted *His:AtSUM1-GG*. The other form of SUMO could not be ligated to protein targets and was denoted *His:AtSUM1-AA*, this form was used as a negative control in the SUMOylation assay. The plasmids were transformed into *E. coli* BL21 (DE3) strain through repeated rounds transformation and chemically competent steps. A third plasmid containing the gene of interest was transformed into the system. This third plasmid was transformed into both the positive and negative control types of cell (*His:AtSUM1-GG* and *His:AtSUM1*-AA). The pDESTxx plasmids were used in this assay and triple antibiotic selection using carbenicillin, spectinomycin and chloramphenicol was used to maintain the plasmids.

SUMOylation assays were performed by first growing 10 ml starter cultures inoculated from a single bacterial colony. These cultures were incubated overnight at 37°C with shaking. These cultures were then used to inoculate 20 ml Erlenmeyer flasks containing 20 ml of Luria-Bertani LB media using 1 ml of starter culture. Cultures were then grown at 37°C for 30 minutes and then chilled briefly. Expression was induced by the addition of 0.1 mM IPTG. Cultures were then incubated at 28°C for 2 hours and then bacterial cells were collected for protein analysis.

2.3.12 *In vitro* cell free SUMOylation assay of RGA

A cell free assay using E1 and E2 enzymes to SUMOylate RGA *in vitro* was used. The assay was adapted from the method by Park-Sarge & Sarge (2009) and lacked the ATP regeneration mechanism used in the original method.

Required reagents

- Storage buffer (50 mM Tris, pH 7.6)
- 10x reaction buffer (500 mM Tris, 500 mM KCL, 50 mM $MgCl_2$, 10 mM DTT, pH 7.4)
- 10 mM ATP solution, pH 7.0
- Inorganic pyrophosphatase solution

Protein expression

Expression vectors for AtSUM1 and an E1 dimer consisting of a dual expression vector containing *His:SAE1a* and *S:SAE2* were used from the reconstituted system from Okada *et al.* (2009). The E2 enzyme, *SCE1*, was cloned from *Arabidopsis* cDNA into pDEST17 to create an N-terminal His fusion. C-terminal His tagged RGA was expressed from *RGA* int pET DEST 55. E1, E2 and RGA vectors were expressed in BL21 CodonPlus (DE3) RIL *E. coli* cells (Agilent Technologies) while the His:SUM1 was expressed in BL21 (DE3) cells. Expression cultures of 250 ml of LB media was used for the E1 and E2 vectors while for the *His:AtSUM1* and *RGA:His* vectors, 500 ml was used. The cultures were inoculated from starter cultures at a ratio of 1:20 and grown at 37°C with vigorous shaking until the OD_{600} of cultures reached 1.0. The cultures were then briefly chilled on ice for 5 minutes to lower their temperature to below 30°C then IPTG was added to a final concentration of 0.1 mM to induce recombinant gene expression and were then grown at 30°C with vigorous shaking for 2 hours. Proteins were then purified as per the method for Ni-NTA HisTrap purification described earlier. The E1 was purified as an intact complex with S:SAE2 co-purifying with His:SAE1a.

Once the proteins had been purified, they were concentrated and buffer exchanged into storage buffer using 0.5 ml Amicon centrifugal spin columns. RGA:His and His:AtSUM1 were concentrated to 1 mg/ml, flash frozen in liquid nitrogen and stored at -80°C until use. The E1 and E2 enzymes were concentrated to 0.1 mg/ml and then mixed with an equal volume of glycerol to form 50% solutions with a concentration of 0.05 mg/ml and these were stored at -20°C.

Assay

SUMOylation reactions in 20 µl were set up as in Table 2.15 below:

Reagent	Volume (µl)	Amount/concentration in reaction
E1	1	50 ng
E2	1	50 ng
RGA:His	5	5 µg
His:AtSUM1	5	5 µg
Pyrophosphatase (0.1 U/µl)	1	0.01 U
10x Reaction buffer	2	1x
10 mM ATP	2	1 mM
H_2O	3	-

Table 2.15: Cell free *in vitro* SUMOylation assay reaction components.

Four reactions were set up in 0.2 ml PCR tubes, with three lacking either one of RGA:His, His:SUM1 or ATP as controls. The reactions mixed using a vortex mixer, then briefly centrifuged to collect the contents at the bottom of the tube. The reaction tubes were then incubated on a heat block at 37°C for 2 hours and the 7 µl 4x SDS PAGE sample buffer was added to tubes and they were heated to 95°C for 2 minutes to halt the reaction and denature the proteins. The samples were then split and loaded onto two 10% SDS-PAGE gels for analysis by western blot. One western blot was probed with αRGA (1° 1:10 000; 2° 1:20 000) and the other with αAtSUM (1° 1:20 000; 2° 1:30 000) for two hours and then processed as per protocol discussed previously.

2.4 Peptide array methods

2.4.1 Large-scale cellulose peptide array screen

Peptide array manufacture

The first monomer of each peptide in the array was immobilised to cellulose sheets and then fluorenylmethoxy-carbonyl chemistry was used to extend the polymer chains using the SPOT peptide synthesis method described in Bolger *et al.* (2006). (Note: synthesis of peptide arrays was not carried out by myself but by a collaborating group at the University of Glasgow).

His:AtSUM1 protein expression

To prepare the *Arabidopsis* SUM1 protein, *AtSUM1* with a hexahistidine N-terminal fusion in the plasmid pCDFDuet was obtained from Okada *et al.* (2009) for expression (*His:AtSUM1*). The plasmid was transformed into *E. coli* strain BL21 (DE3) and successful transformants were selected on LB agar supplemented with 50 µg/ml spectinomycin. Protein expression was performed in 500 ml of LB and expression was carried out at 37°C according to the protein expression protocol. Protein was purified according to the purification protocol for His tagged proteins using the ÄKTA FPLC system. After purification buffer exchange into a 50 mM Tris pH 8.0 solution and protein concentration was performed using a 500 µl Amicon selectively permeable membrane column with a 3000 kDa molecular weight

cutoff and the final concentration of the protein was adjusted to 1 mg/ml.

Large-scale array far-western blotting

Peptide arrays were wetted by washing in 100% ethanol, rinsing in TBST 3 times then equilibrating in TBST for 10 minutes on a rocking platform. The arrays were then blocked with blocking buffer (TBST + 5% w/v Marvel milk powder) in a rotating glass hybridisation cylinder for 2 hours at room temperature. The arrays were then rinsed with 20 ml of TBST and then immersed in 20 ml of probing buffer (TBST + 1%w/v milk) with SUMO protein (0.2 μg/ml for His:AtSUM1 2.0 μg/ml for GST:HsSUM1) and were incubated overnight in a cold room (0 - 5°C) in a rotating hybridisation cylinder. The arrays were then rinsed 3 times with TBST then washed 3 times in TBST for 10 minutes. Primary antibody in 20 ml of TBST was then added. For AtSUM1 polyclonal αAtSUM1 antibody at a concentration of 1:20000 was used. For HsSUM1 polyclonal αGST antibody at a concentration of 1:7500 in 1% milk TBST was used. The primary antibody was incubated for 2 hours at room temperature, after which the arrays were washed in TBST as before. Secondary antibody in 20 ml TBST was then added. For His:AtSUM1, α-rabbit horse radish peroxidase conjugate (HRP) at a concentration of 1:30000 was used. For GST:HsSUM1 α-mouse HRP at a concentration of 1:20000 was used. The concentrations of the primary and secondary antibodies were determined empirically give a similar signal strength between the two different SUMO isoforms. The arrays were incubated in the secondary antibody for 1 hour at room temperature then washed in TBST as before.The arrays were then immersed in 5 ml of enhanced chemiluminescence solution for 20 seconds and then exposed to X-ray film for 2.5 minutes, which was then developed.

The arrays were then stripped to remove interacting proteins and antibodies (see next subsection) and a negative control far-western blot was performed. The same protocol as above was used except that the SUMO protein was left out of the probing buffer. For the GST:HsSUM1 negative control, GST protein was added to a concentration of 1.0 μg/ml to the probing solution. The blots were also exposed to the X-ray film for 5 minutes, twice as long as SUMO protein probed blots. This was to ensure that any non-specific interactions were shown clearly. The negative control blots were used to remove non-specific interacting peptides during data processing.

2.4.2 Cellulose array stripping

The SIM arrays were stripped to prepare them for different probings. Stripping buffer (20 mM dithiothreitol, 60 mM Tris, pH 6.8, 2% sodium dodecyl sulphate) was heated to 70°C in a microwave and then the arrays were placed in the buffer and incubated for 30 minutes with gentle agitation over a hot plate to maintain temperature. The arrays were then removed and washed thoroughly with distilled water, then washed twice for 10 minutes in TBST, then washed in distilled water once more and dried. For

long-term storage the arrays were kept in a -20°C freezer.

2.4.3 Cellulose array Ponceau-S staining

Peptide amounts in the arrays were estimated by staining with the dye Ponceau-S. Due to the short peptide length, Ponceau-S staining does not correlate perfectly with protein amounts. This is due to varying affinity of different amino acid groups to the Ponceau-S dye, therefore this method is only semi-quantitative. The arrays were placed in stain solution (0.1% w/v Ponceau-S, 1% v/v acetic acid) for 30 minutes with agitation, then throughly rinsed in distilled water then washed twice more in water for 10 minutes. The arrays were then dried and photographed. The Ponceau-S stain was removed by washing the arrays in 200 ml of distilled water with 2-3 drops of 5 M NaOH twice for 10 minutes and then stripped using the previous stripping protocol.

2.5 Small-scale nitrocellulose peptide array screen

Synthetic peptides were manufactured by Cambridge Research Biochemicals for use in the nitrocellulose arrays. Solutions of 1 mg/ml of peptide were prepared and 1 µl of the peptides were spotted onto a nitrocellulose membrane to make the array. The array was left to stand for 5 minutes and was then placed in a petri dish and rinsed with TBST. The array blocked with TBST with 5% w/v Marvel Milk power for 5 minutes on a rocking platform. AtSUMO was then added to the blocking buffer to a final concentration of 0.05 µg/ml. The array was then incubated for 1 hour. The array was then washed three times in TBST for 5 minutes. Primary antibody (αAtSUM1, 1:30000) in TBST was incubated for 1 hour. The array was washed then the secondary antibody (αRb HRP, 1:20000) was incubated for 1 hour then washed again. The array was then developed as per the western blotting protocol.

2.6 Plant methods

2.6.1 Seed sterilisation

The chlorine vapour phase surface sterilisation method was used. Seeds were placed in open 1.5 ml tubes labelled with pencil and placed in a sealable plastic box in a fume hood. A flask containing 100 ml 10% sodium hypochlorite solution was placed in the plastic box then 3 ml of 37% HCl was added and the plastic box promptly sealed. The seeds were left overnight for surface sterilisation to occur. The next day, the box was opened and the flask removed and then closed promptly. The plastic box was then moved to a laminar flow hood and the seeds removed.

2.6.2 Floral dip transformation

Arabidopsis was transformed following an adapted version of the original floral dip method by Clough & Bent (1998). *Arabidopsis* plants were grown in pots with 5 plants. For each construct, 4 pots of plants were used. Plants were grown until flowers had emerged and the floral stems were cut off to promote the development of multiple stems. Plants were then used for floral dipping once large floral stems had grown.

Agrobacterium clones with a T-DNA plasmid containing the transformation construct were used to inoculate 10x LB starter cultures with appropriate antibiotics and grown overnight at 28°C with shaking. The starter cultures were used to inoculate 250 ml LB cultures with antibiotics in 1 l Erlenmeyer flasks and these were grown overnight under the same conditions. The following day the *Agrobacterium* cells were pelleted by centrifuging at 5000 g for 10 minutes. The pellets were then re-suspended in 500 ml of 5% sucrose solution then 100 µl of Silwet L77 was added. Prior to transformation, all siliques were removed from the *Arabidopsis* plants. The stems of the plants were then submerged in the *Agrobacterium* solutions for 20 seconds ensuring that all the flowers were submerged. The plants were then placed flat in plastic bags and left for 24 hours. The plants were then removed from the plastic bags and set upright and normal plant growth conditions were resumed. Seeds were collected from these plants and were then screened for transgenic lines.

2.6.3 Germination assay

To perform the germination assay, square 15 cm petri dishes with ½ MS agar were prepared. The agar plates were supplemented with 0, 0.1 or 0.5 µM paclobutrazol (PAC). Seeds for the lines to be tested were surface sterilised and then seeds for each line were spread over three plates with the different PAC concentrations. Around 100 to 250 seeds were applied to each plate. Plates were then placed in a 0-5°C cold room for 48 hours. Plates were then moved to an incubator set to long day conditions. After 48 hours, the germination of the seeds was recorded. Seeds were viewed under a dissecting microscope. A seed was determined to have germinated if both the seed coat had opened and the primary root had emerged.

2.7 Computational methods

The R programming environment version 3.1.1 (R Core Team, 2013) was used for all data processing and statistical analyses. The R programming language was used to develop random forest predictors. MATLAB R2013b (version 8.2) was used to develop image analysis software. Full theoretical methods for computational work are described in Chapter 3.

Chapter 3

Prediction of SUMO-related sequence features

3.1 Introduction

SUMO plays an important role in cellular function and is implicated in a vast number of molecular processes (Hay, 2005). With the wealth of genomic data now available, computational methods to predict SUMOylated lysine residues (referred to as 'SUMO sites' hereafter) and SUMO interacting motifs (SIMs) are becoming evermore important. Predication of possible sequence features from primary sequence data is used both to identify new research targets and to identify the location of sequence features in proteins that are known to have such features. SUMO sites and SIMs will be collectively referred to as 'SUMO sequence features' in this chapter.

A number of methods have been developed to predict SUMO-related sequence features and tools to predict SUMO sites were the first to emerge. Currently there are three methods with online interfaces, two of which are published. SUMOplot was the first publicly available SUMO site predictor to be released. It was developed by Abgent but the method was not published so the strategy used to predict SUMO sites is not known. Later SUMOsp version 1 was released, which was the first published method and was developed by Xue *et al.* (2006). SUMOsp compared a query sequence with known SUMO sites using the sum of substitution scores from the BLOSUM62 matrix. BLOSUM62 is a protein substitution matrix developed to score amino acid similarities for protein sequence alignment (Henikoff & Henikoff, 1992), though it is often used in protein similarity scoring as well. SUMOsp was later updated to version 2 which used a novel scoring matrix that was optimised using a genetic algorithm. It also introduced partitioning of the results into two categories of SUMO site, types I and II (Ren *et al.*, 2009). Type I sites conform to the canonical SUMO site motif of ΨKx[ED], while type II sites do not completely conform and the prediction of these sites is less accurate. There is however, no functional distinction between the two types. The methods mentioned so far all had one thing in common: they all used amino acid factors (i.e. letters to represent the amino acid groups) in the prediction and did not directly include any numeric variables of amino acid chemical properties in the prediction method.

The most recent SUMO site predictor, seeSUMO, took steps to address this lack of inclusion of chemical data (Teng *et al.*, 2012) by incorporating numerous numeric amino acid properties. seeSUMO uses data from the AAindex, a database of hundreds of indices of amino acid properties (Kawashima & Kanehisa, 2000). Teng *et al.* (2012) converted the amino acid factor levels into vectors of numeric properties to capture the chemical information of the amino acids in order to build a random forest classifier to predict SUMO sites. Converting amino acids into numeric vectors of chemical properties and using the more powerful random forest classifier led to the development of a predictor that outperformed previous tools. However, the performance of seeSUMO was far from optimal as there were a number of critical drawbacks in the methodology used byTeng *et al.* (2012) that negatively impacted the predictor. The authors did not perform parameter reduction and the top performing random forest predictor used an input of 20 amino acids, each of which was converted into a vector of 40 numeric properties. It is

unclear whether the authors removed the central lysine from the input so the parameter space was either 760 or 800 dimensions. A large number of these parameters were highly correlated and this introduced a large amount of redundant information into the model.

A predictor with such a high dimensionality will be subject to the Hughes phenomenon whereby predictor performance declines if more parameter dimensions are added after the optimal number (Hughes, 1968). One of the factors that contributes to the Hughes phenomenon is that as the dimensionality of parameter space grows, the hyper-dimensional volume of that parameter space expands exponentially. Training data projected into this large hyper-volume then becomes sparse, with similar data-points becoming ever more distant. To counteract this sparsity, the training data would need to grow exponentially with the number of dimensions, something that is typically not feasible from a practical perspective. The Hughes phenomenon suggests that there is an optimal dimensionality, where the number of dimensions is high enough to effectively compartmentalise the different data classes but not so high that data sparsity becomes problematic. Therefore parameter selection and model simplification should always be implemented to improve predictor performance.

The seeSUMO predictor also partitioned training data into training and evaluation sets which is generally not necessary with random forest models. Random forests, if used correctly, can calculate an unbiased (if not pessimistic) estimate of error from the training data alone using out of the bag (OOB) error estimation, negating the need to have a separate evaluation set for predictor performance assessment (Breiman, 2001). Using all of the possible training data may have improved the performance of the predictor.

While there are a number of SUMO site predictors available, prediction of SIMs in proteins is still a very new area. Currently there is only one publicly available predictor, which was developed by the same group that developed SUMOsp, but the method has not been published and importantly does not take SUMO isoform differences into account. One difference between SUMO sites and SIMs is the amount of available data for training predictors. Hundreds of SUMO sites have been published and this has allowed researchers to build training sets of more than 1000 test sequences (these data are mostly negative, non-SUMO site sequences). The same is not true for SIMs; while a large number of proteins have been published that interact with SUMO, only in a small number of these publications is the site of interaction determined, so there is a desperate need for novel training data. Another interesting aspect of SIMs is that different SUMO isoforms within a species show SIM motif specificity, Namanja *et al.* (2012) investigated the human PIASx SIM in detail and found that mutagenised versions of this SIM could be made to interact specifically with either HsSUM1 or HsSUM3 or with both. These results showed distinct divergence in SIM sequence and SUMO isoform interactions. This is in contrast to SUMO sites where there has been no observed influence of motif composition on the SUMO isoform ligated. Also, it is generally thought that there are no major species differences in SUMO site composition and SUMO

site predictions are assumed to be applicable across species.

Since there are known SUMO isoform preferences for SIM sequences within species, it is highly likely that there will be interspecies differences too. A predictor trained on data from one species may not be applicable to a different species, therefore a SIM predictor will need to be designed to specifically predict SIM binding to different SUMO isoforms and to homologous SUMO proteins in other species.

3.1.1 SIM features

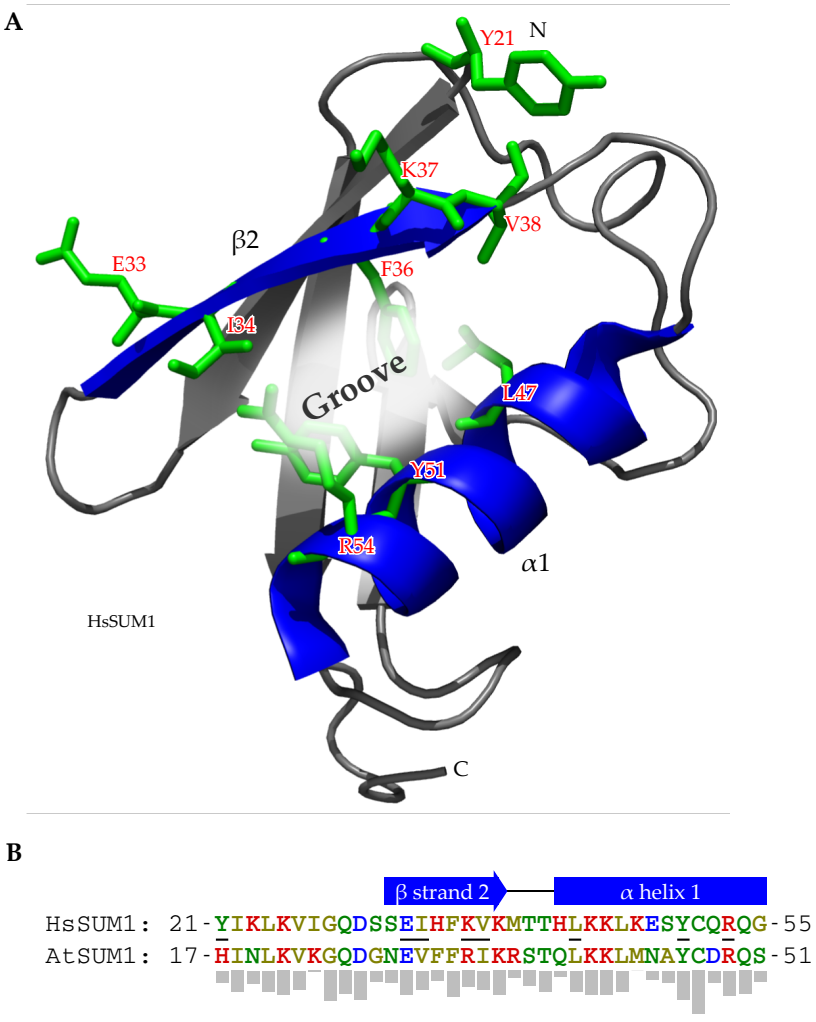


Figure 3.1: SIM binding site in SUMO. **(A)** 3D structure of HsSUM1. SIMs bind to amino acids in the groove between β-strand 2 and α-helix 1. The side chains of amino acids that have been shown to make contact with SIM amino acids are shown in green. **(B)** Alignment of the HsSUM1 and AtSUM1 sequences involved in SIM binding, bars between the alignment correspond to the highlighted side groups shown in the 3D structure. While there is a large amount of sequence divergence between the two species, amino acids responsible for SIM binding are conserved. The bars below the amino acids represent the conservation score. Colours of the amino acids represent the following properties: yellow: hydrophobic, green: polar, blue: negative charge and red: positive charge. 3D structure in this figure was solved using NMR by Song *et al.* (2005).

SIMs are short peptide sequences within a protein that bind to SUMO through non-covalent interactions and these interactions play an important role in the assembly of protein complexes. Initially

there had been some disagreement as to which parts of a SIM were responsible for the interaction with early researchers incorrectly suggesting that the interaction was due to and SxS motif (Minty *et al.*, 2000). The actual sequence responsible for SUMO interaction was solved by nuclear magnetic resonance (NMR) studies of the human RanBP2 SIM by Song *et al.* (2004) who showed that a short stretch of hydrophobic amino acids fits into, and binds to, a groove within SUMO. This groove within SUMO is flanked by an α-helix and a β-strand (Figure 3.1) and the side groups of these secondary structures make contact with SIM side chain groups to form the interaction. When SIMs are not bound to SUMO they are almost always unstructured polypeptide chains, but once this chain binds to the SUMO groove it takes on a structured conformation of β-strand that extends the β-sheet within SUMO (Namanja *et al.*, 2012). SIMs have the rather unique property that they can bind to the SUMO groove in either peptide chain direction, either parallel to or antiparallel to the second β-strand within SUMO (Song *et al.*, 2005) and some SIMs exist that can bind in both directions, though usually a particular SIM tends to bind in the orientation with the lowest binding free energy.

Further research showed that there are three distinct classes of SIM, termed A, B and R (for reverse), though the overall structure with a hydrophobic core is consistent between these motifs (Vogt & Hofmann, 2012). The most common amino acids found within the hydrophobic cores are leucine, isoleucine and valine though other hydrophobic amino acids can also be substituted into the core. SIM type A has a general motif structure of ΨΨxΨ and is best suited to extending the SUMO β-sheet in an antiparallel orientation while the reverse of this SIM, type R has a motif structure of ΨxΨΨ and extends the SUMO β-sheet in the parallel orientation (Namanja *et al.*, 2012). The final type of SIM, type B, is slightly different from the other two and has a core motif structure of ΨΨDLT and, like SIM A, usually extends the SUMO β-sheet in the antiparallel orientation.

While the hydrophobic core of the SIM is responsible for the majority of the interaction with SUMO, negatively charged sequences flanking the core also interact with SUMO and stabilise the conformation of the two proteins. The flanking residues also determine SUMO isoform selection by promoting a stronger interaction with a particular SUMO isoform (Hecker *et al.*, 2006). SIMs often show specificity, or at least higher affinity, for a particular SUMO isoform and this is partly responsible for the functional diversification of different SUMO isoforms (Zhu *et al.*, 2009). Since these flanking regions are responsible for isoform selection, they offer the ability to predict which SUMO isoform a particular peptide is most likely to bind to and suggests that training data used to build a predictor for a particular SUMO isoform will not be applicable to predicting other isoforms.

Post-translational modification (PTM) plays an important role in regulating SIM-SUMO binding and can act as a switch to rapidly modulate the behaviour of SUMO. Often SIMs contain serine residues flanking their hydrophobic core and phosphorylation of these polar residues imparts a negative charge on them that promotes interaction with SUMO. SIM phosphorylation has been found to both enhance

interaction strength with SUMO (Mascle *et al.*, 2013) and to switch the interaction between two SUMO isoforms (Hecker *et al.*, 2006). Acetylation of SUMO is another PTM that has been found to regulate interaction. In HsSUM1, the positive charge of K37 is required for SIM interaction and acetylation of this group neutralises and blocks the interaction with SIMs. Interestingly acetylation of SUMO promotes interaction with bromodomain-containing proteins (Ullmann *et al.*, 2012), which are typically transcriptional regulators (Mazur & van den Burg, 2012), thus acetylation acts to switch SUMO interaction between SIMs and bromodomain containing proteins.

3.1.2 SUMO site features

Covalent attachment of SUMO occurs at lysine residues within a target protein. These SUMOylatable lysines often conform to motif ΨKx[ED] known as a SUMO site, which is conserved across eukaryotes (Hay, 2005). The function of the SUMO site is to attract an activated SUMO-E2 thioester linked complex and promote conjugation of SUMO to the lysine residue. The motif has been shown to be important in facilitating SUMO conjugation but about 20% of known SUMO sites do not conform to this motif. However, these non-consensus SUMO sites have been found to have lower conjugation efficiencies (Schwamborn *et al.*, 2008) and so are likely to be weaker targets of SUMOylation *in vivo.* Alternatively, SUMOylation at type II sites could be facilitated by some mechanism other than attraction to the core motif. More recent research into the composition of SUMO sites has revealed complexity beyond the core ΨKx[ED] motif. The amino acid positions immediately upstream of the motif are often enriched for negatively charged amino acids. The role of this charged site is to facilitate interaction with the E2 part of a SUMO-E2 complex during conjugation, as the E2 enzyme contains a corresponding positively charged patch that associates with this negatively charged site in the SUMO site motif (Yang *et al.*, 2006).

The amino acids in the variable parts of the ΨKx[ED] motif have significant effects on the efficiency of SUMOylation with aromatics at the large hydrophobic position (Ψ) greatly enhancing the conjugation rate. Glycine, proline or negatively charged amino acids at the third position (x) reduce SUMO conjugation (Schwamborn *et al.*, 2008); for proline, the steric hindrance imposed by this amino acid may lead to less favourable conjugation conditions. The emerging understanding is that the SUMO site motif is more complex than the originally published ΨKx[ED] motif often used in the literature to summarise SUMO sites and using more complex amino acid features can allow efficient prediction of SUMO sites. Phosphorylation of the SUMO sites has been found to act as a SUMO conjugation switch with the serine in the motif ΨKxExxSP being the target of phosphorylation which imparts an enhanced negative charge to the SUMO site and promotes interaction with the E2 via its positively charged patch (Mohideen *et al.*, 2009). Phosphorylation of SUMO sites however, was not investigated in this research project.

Apart from the interaction between the SUMO-E2 complex and a target SUMO site, E3 ligases enhance SUMO conjugation analogous to the E3s in the ubiquitination cascade. However, unlike the ubiquitin E3 ligases, SUMO E3 ligases are not necessary and conjugation will still occur, albeit at a reduced rate. SUMO E3s have been shown not to recognise the target protein but rather to orientate and restrict the flexibility of the thioester bond between SUMO and the E2 enzyme (Reverter & Lima, 2006) which reduces the activation energy of SUMO conjugation (Truong *et al.*, 2011). Since the E3 enzymes do not directly detect target proteins, SUMO conjugation is strongly governed by the sequence surrounding the target lysine and not detection of target protein specific features by the E3 as is the case with ubiquitin E3 complexes (Sadanandom *et al.*, 2012), which makes SUMO sites good candidates for prediction.

3.1.3 Random forest classifiers

Random forests were developed by Breiman (2001) and are an ensemble machine learning method used for classification or regression of data. For the benefit of simplicity, only the classification aspect will be considered in this chapter. Data to be classified consist of N observations with M variables and random forests are capable of handling large numbers of variables which can either be continuous or nominal (alternatively referred to as factors in statistics). Ensemble methods consist of many small objects that perform a very simple classification task and the output from these many small objects is then combined to give a single result for the collection. In the case of random forests, the object is a decision tree with nodes that split into exactly two branches. At each node, an input observation is split down the tree according to the value of some of the input variables. At each node a set of m variables is used to determine the split, with the variables and direction of the split decided during the training of the tree. Once an observation reaches a terminal node on a decision tree, a vote for a particular class is made for the observation.

Overall the performance of decision trees is poor. However, building an ensemble of trees and then averaging the votes of all the tree results in significantly improved performance and the method has a robust built in error estimation mechanism eliminating the need for a validation set during classifier training. While some classifiers have many parameters that need to be optimised during training, such as vector support machines, m is the only parameter that needs to optimised within random forest models, though in practice variable selection is generally required especially when there are a large number of input variables. Increasing the value of m up to the point $m < M$ results in better performance of any particular tree but also increases the correlation between trees in a random forest which causes a decline in the overall random forest performance hence an optimal value of m should be found for best performance. In practice the value of m is usually much smaller than the total number of variables ($m \ll M$).

During building of each tree in the random forest, N observations from the training data are sampled with replacement, giving each tree a random bootstrap of the training data. Two thirds of this bootstrap are then used to build the decision tree, randomly choosing m variables at each node to test. The remaining third of the bootstrap or the out of bag (OOB) data is then tested on the newly built decision tree to assess its performance. This process is repeated until the desired number of trees are built and a global estimate of error can be calculated from the combined OOB data. Because the error for each tree is estimated using data that was not used to build it, the tree error estimate is unbiased, thus so is the global error estimate of the random forest (Breiman, 1996). By bootstrapping and partitioning data separately for each tree, all data is used for training the random forest, a particular data point on average will be used in the construction of about two thirds of the decision trees. Using this per tree bootstrap method allows all of the training data to be used to train a random forest and allows an unbiased estimate of predictor error to be calculated and eliminates the need to partition the training data into training and evaluation sets. This allows the random forest to fully utilise all training data. This is especially useful when there is limited training data.

The error estimate used internally by the random forest algorithm is the proportion of times data from the classification with the most positive votes is not equal to the correct class, or the false positive rate (FPR) of the classification with the most votes. Other classifier performance assessments can also be calculated using the OOB results such as sensitivity (true positive rate; TPR) or a receiver operator characteristic (ROC) curve. The sensitivity and the FPR of a predictor depend on the value of the predictor cutoff, which in the case of the random forest is the value of the proportion of votes at which an input is classified as positive. By default this cutoff is set at 0.5, that is to say an input needs to get a positive vote from at least half of the trees in the random forest. Increasing the cutoff value reduces the FPR of the random forest predictor but also decreases the sensitivity, thus the FPR or sensitivity can be tuned to a specific value by varying the cutoff value. Because the values of the FPR and sensitivity are inversely dependent on each other, either by itself is not a useful assessment of the performance of a predictor without taking the other into account. The ROC curve on the other hand is a plot of the FPR against the sensitivity for all values of the cutoff and gives an overview of all possible FPR and sensitivity values achievable with a given predictor.

In a perfect predictor, the ROC curve occupies upper left bounds of the graph while a predictor which is no better than random performance has a curve that runs diagonally across the graph. The area under the ROC curve (AUC) can thus be used as a metric to assess the performance of a predictor independent of the cutoff value and is a very useful metric to reliably compare different predictors. A perfect predictor has an AUC value of 1.0 while a predictor which performs no better than random has a value of 0.5. From a statistical perspective, the AUC value is the probability that a classifier will score a randomly chosen true positive instance higher than a randomly chosen true negative instance (Hanley

& McNeil, 1982). The AUC value is related to another performance metric, the Gini coefficient, which is twice the area of the upper diagonal of a ROC curve or $2 \times \text{AUC} - 1$ (Fawcett, 2006). The Gini coefficient is useful as it remaps the useful AUC interval of [0.5, 1] to [0, 1], giving a more intuitive range of values.

The Gini coefficient is used internally by the random forest algorithm to score the importance of the input variables and this feature can be used to identify which variables to exclude from the predictor to reduce its complexity.

3.1.4 Improving current models

The motivation of the work described in this chapter was to build a predictor of SUMO sites and SIMs by further developing the method of using random forest predictors. The SUMO site predictor was trained using data from Ren *et al.* (2009). As there were insufficient examples of SIMs in the literature, especially for plant models, data was generated from a library of synthetic SIM-like peptides. Predictors specific to particular SUMO isoforms were built to address the issue of SUMO isoform binding differences. To improve upon previous SUMO feature predictors, more robust variables were used to represent the peptide sequences and rigorous variable selection was performed with the goal of minimising information redundancy and model complexity in order to optimise model performance.

The predictors were used to build a graphical web-based tool that can be deployed for use to the wider research community. The tool includes a new feature not seen in current SUMO feature models; results are overlaid onto a multiple sequence alignment of the input sequences. This allows straightforward identification of conserved SUMO sequence features, which have a higher probability of being true functional features.

Finally the completed SIM predictor was used to search for conserved predicted SIMs in orthologous proteins from multiple plant species. These predicted SUMO binding proteins (SBPs) were then analysed to identify enrichment in any biological and biochemical functions and constitutes the first large scale analysis of SBPs in any organism.

3.2 Chapter aims

- Generate a peptide library to screen for SUMO1 interacting peptides.
- Characterise and compare human and plant SIMs.
- Use generated SIM data and published SUMO site data to build random forest models to predict these features.

- Incorporate amino acid chemical features from the AAindex database, and evolutionary information into the models.
- Screen the *Arabidopsis* proteome for predicted SIMs and analyse the results.
- Develop a web-based user interface for the sequence feature predictor.

3.3 Materials and methods

3.3.1 SIM peptide array design

To generate the SIM data to be used to train random forest predictors and for feature analysis, a library of synthetic SIM-like sequences was created. 11-mer SIM-like peptide sequences were designed, using published SIM motifs from Vogt & Hofmann (2012) as a template. These motifs were generated from the limited SIM data collected from animal and yeast research and were not specific to any SUMO isoform. The published SIM motifs were used to generate sequence models using regular expressions for each SIM type, A, B and R. The regular expressions encoded which amino acids were allowable at a particular position along the 11-mer peptide. The core regions of the SIM models only permitted hydrophobic amino acids for example. For each SIM type there was a stringent regular expression model, conforming strongly to the published motif, and a number of more variable expressions allowing more amino acid sequence diversity. The models, written as Perl style regular expressions, are shown in Table 3.1. The purpose of these SIM models with varying stringency was to generate a set of peptides with a large amount of sequence variety and to sample peptides with evermore divergence from the published motifs.

Class	Perl style class model
A1	`..[IVLMP][IVLM].[IVLM][^VLIMFYWAP][^VLIMFYWAP][^VLIMFYWAP]....`
A2	`..[IVLMP][IVLM].[IVLM][DSE][DSE][DSE]....`
A3	`..[IVLMPFYW][IVLMFYWP].[IVLMFYWP][DSE][DSE][DSE]....`
B1	`..[^GDN][IVY]DL[TY]......`
B2	`..[^GDN][IVYLMFWCAP]DL[TY]......`
B3	`..[^GDN][IVY]DL[TYDEFSC]......`
B4	`..[^GDN][IVYLMFWCAP]DL[TYDEFSC]......`
R1	`....[^VLIMFYWAP][^VLIMFYWAP][^VLIMFYWAP][IVLM].[IVLM][IVLM]..`
R2	`....[DSE][DSE][DSE][IVLM].[IVLM][IVLM]..`
R3	`....[DSE][DSE][DSE][IVLMPFYW].[IVLMPFYW][IVLMPFYW]..`

Table 3.1: SIM class models used to search the *Arabidopsis* proteome for test sequences. Models are encoded with Perl style regular expressions. Full stop (.) = any amino acid. Square brackets ([ ]) = any one of those amino acids inside brackets. Caret (^) inside square brackets = any amino acids except those inside the brackets.

To generate sequences from the SIM models, the models were used to search translated representative *Arabidopsis* gene sequences (from the TAIR10 genome assembly; translated representative gene models). Sequences were searched against the *Arabidopsis* translated genome to ensure that natural amino acid frequencies were sampled and that any known or unknown natural sequence restrictions

were adhered to. There was concern that using a stochastic model to randomly generate sequences would likely violate underlying sequence composition restrictions. Duplicate sequences were removed and then sequences were randomly selected for inclusion in the SIM library. For each of the three SIM types, around 200 peptides were chosen. 8 control peptides were included of known interactors or non-interactors and in total the library contained 600 peptides.

To investigate the role of phosphorylation on the binding of SIMs to SUMO, a second library of phosphorylated SIM-like peptides was designed. The amino acids serine, threonine and tyrosine are all targets for phosphorylation *in vivo* and SIM peptides with phosphorylated versions of all three of these amino acids were designed. The phosphorylated peptides were not used to train the SUMO predictor models, rather they were used to identify important phosphorylation sites in SIMs. For this library, template non-phosphorylated peptides were designed using the same strategy as the first library, however only one SIM model per class was used, the model generating the highest number of positive interactions, the models A1, B4 and R1. These models were identified by performing a preliminary screen of the first array to calculate the rates of positive interactions from peptides designed using the various SIM models. Template peptides without any phosphorylations were include in this array to act as negative controls. The template peptides were also selected so that they were unique from any peptides in the first library; this allowed integrating these template peptides into the dataset from the first library, allowing more peptides to be used for training the SIM predictors.

To generate the phosphorylated peptides, the template SIMs were used to design multiple phosphorylated versions of the template. Once this was complete, the template peptide and the phosphorylated versions were added to the second library. Eight control sequences were added to the array and a total of 600 peptides were included. The pseudocode for generating the phosphorylated SIMs is shown in Algorithm 1.

Once the peptide libraries were designed, data were prepared for synthesis on 30 x 20 spot arrays. Control sequences were placed on their respective arrays in a pattern that allowed unambiguous identification of array orientation. For the non-phosphorylated array, the order of the remaining peptides was completely randomised to accommodate for any possible array positional effects in peptide synthesis quality. The peptides in the phosphorylation array were grouped by template peptide and then the order of the groups was randomised. The peptide arrays were synthesised using an automated system.

3.3.2 Interaction screen of SIM peptide arrays

The peptide arrays were synthesised using an automated system. To identify peptides with affinity for SUMO, far-western blotting using *Arabidopsis* SUM1 (AtSUM1) and human SUM1 (HsSUM1) was performed. Recombinant HsSUM1 fused to glutathione *S*-transferase (GST:HsSUM1) was purchased from Enzo life sciences (product code:UW0150) and AtSUM1 was prepared using recombinant expres-

Algorithm 1 Pseudocode for phosphorylated SIM generating algorithm

For each SIM model, find all matching peptides in translated *Arabidopsis* genome

Remove all duplicate sequences

Randomise order of peptides

for *Each SIM* **do**

  if *SIM has no S, T or Y groups* **then**
    Discard SIM and choose next SIM
  end
  if *SIM has 1 or 2 phosphorylatable amino acids* **then**
    Generate all possible phosphorylation combinations
  end
  if *SIM has 3 or more phosphorylatable amino acids* **then**
    Randomly generate 2 SIMs with 1 phosphorylation
    and
    Randomly generate 2 SIMs with 2 phosphorylations
    and
    Randomly generate 1 SIM with 3 phosphorylations
  end
  Add peptide sequences to library

end

sion in *Escherichia coli* (*E. coli*). Interaction with peptides in the arrays was screened using far-western blotting. The quantity of peptide in the arrays was estimated by staining with Ponceau-S. See Chapter 2 for full molecular biology methods. Due to the peptide synthesiser being used by multiple researchers and synthesis runs being queued, the unmodified peptide library split over in two parts.

3.3.3 SIM array image data collection

The far-western film images and the dried Ponceau-S stained arrays were imaged with a Nikon digital camera to collect peptide spot intensity data. This section describes the methods used to measure and normalise the intensity data from the images of the SIM peptide arrays. As the far-western images represent interaction data, they will be referred to as simply the interaction images hereafter. The Ponceau-S stained images on the other had show total protein levels in the arrays and will be referred to as the protein quantity images hereafter.

3.3.3.1 Image preprocessing

All preprocessing work was carried out in the image processing software Fiji (Schindelin *et al.*, 2012), a distribution of ImageJ for biological image processing. The colour images were converted to tagged image file format (TIFF) and then converted to 32 bit greyscale (floating point) images. Artefacts such as dust and speckling were locally removed using a noise filter. The images were then rotated so that the peptide spots aligned with the x and y axes of the image, this was required for correct spot identification in subsequent processing. The images were then scaled by 0.5 in both the x and y direction (75% reduction in area) using bilinear interpolation.

The background of the interaction images (but not the protein quantity images) was then normalised by first generating an approximation image of the background and subtracting this from the original image. To generate the background, the intense regions from the interacting spots were filtered out using a low-pass noise filter and then the resulting image was smoothed using a Gaussian kernel with a radius of 50 pixels. These images, along with the protein quantity images were then reduced to 8 bit greyscale images (256 integer scale values) and were saved for further processing.

3.3.3.2 Intensity data collection

The processed array images were analysed using a MATLAB tool with a graphical user interface written specifically for the purpose (see appendix C.1 for source code). Centre points for each peptide spot were calculated based on the grid size and spacing. Discs of the same size as the peptide spots were then drawn over each spot and the average intensity of the pixels beneath the discs was calculated. To calculate which pixels to include in the mean calculation for a particular peptide spot, any pixel in the image, (X_i, Y_j), had to satisfy the inequality

$$(X_i - x)^2 + (Y_j - y)^2 \leq r^2$$

to be included, where $(x,\ y)$ is the centre of the disc over a peptide spot and r is the radius. The intensity data for each spot was then matched with its corresponding peptide sequence and annotation.

Background data were also collected from the arrays for use in later normalisation steps. The interaction images and quantity images were treated differently in this process as the interaction images had signal that bled into the inter-spot space on the arrays. For the interaction images 10 evenly distributed points in areas of the image with no interaction signal were selected and the mean and standard deviation of the intensities was calculated. For the quantity images, the intensity of an area next to each spot was measured so that each peptide value had a corresponding background value.

3.3.3.3 Data normalisation, scaling and correction

Once the numeric data had been extracted from the array images, all further processing and analysis was performed in R (R Core Team, 2013) and the R package and package ggplot2 (Wickham, 2009) was used to generate figures.

To remove discontinuities between the different parts of the arrays, intensities were normalised against the background values and rescaled. The approach differed for the intensity and quantity array values. For each part of the interaction arrays, the mean background value, μ, was subtracted from the matrix of intensity values, $\mathbf{I}$, and then the data was rescaled between 0 and 1 using the 0th and 97.5th percentile. The 97.5th percentile was used to prevent extreme values distorting the data. The normalisation and rescaling steps are given by:

$$\mathbf{I}_{\text{normalised}} = \mathbf{I}_{\text{original}} - \mu$$
$$\mathbf{I}_{\text{scaled}} = \frac{\mathbf{I}_{\text{normalised}}}{P_{97.5}(\mathbf{I}_{\text{normalised}})}$$

Normalisation and scaling of the quantity image parts was performed slightly differently. Rather than subtracting some global background mean, the background intensities for each spot that had been estimated earlier were subtracted from the peptide intensities, $\mathbf{Q}$. The resulting values were then scaled between 0 and 1, again using the 0th and 97.5th percentile. These steps are given by:

$$\mathbf{Q}_{\text{normalised}} = \mathbf{Q}_{\text{original}} - \mathbf{Q}_{\text{background}}$$
$$\mathbf{Q}_{\text{scaled}} = \frac{\mathbf{Q}_{\text{normalised}}}{P_{97.5}(\mathbf{Q}_{\text{normalised}})}$$

The peptide quantity data were then used to rescale the interaction data, essentially correcting intensity values for spots with low peptide levels or no peptide. Spots with peptide levels close to 0 were removed from the data as these could not be used to extrapolate interaction levels.

3.3.3.4 Error estimation and data exclusion

Correcting the interaction values by dividing by the quantity values introduces error into the final result. The 95% confidence interval for this error, $\hat{\epsilon}$, was estimated as

$$\hat{\epsilon}_i = \frac{1.96\sigma}{\mathbf{Q}_{\text{scaled}_i}}$$

where σ is the calculated standard deviation of the background of the scaled intensity images. A cutoff for $\mathbf{I}_{\text{scaled}}$ by $\hat{\epsilon}$ was chosen since $\hat{\epsilon}$ becomes large for small values of $\mathbf{Q}_{\text{scaled}}$, as

$$\lim_{\mathbf{Q}_{\text{scaled}} \to 0} \frac{1.96\sigma}{\mathbf{Q}_{\text{scaled}}} = \lim_{\mathbf{Q}_{\text{scaled}} \to 0} \hat{\epsilon} = \infty$$

As the error associated with any particular intensity value increases, the uncertainty about the true value of the intensity increases and a suitable error calculation to exclude values was determined. Rather than using a single threshold to decide which data to exclude, the intensity was taken into account as some values with high error may have such a high intensity value that they can still be assumed to be a true interacting spot. An example of this would be a spot with a very low peptide amount that interacts very strongly with SUMO.

As well as the estimated error, the peptide spot intensities from the negative controls had to be included in the criteria for removing data. The intensity of the control data, which was processed in the same manner as the interaction data, is given by $\mathbf{C}$. The *Arabidopsis* and human data were analysed using two different methods due to very high noise in the human data and a high number of non-specific interactions. For the *Arabidopsis* data, three thresholds were defined: maximum negative control intensity threshold ($T_{\mathbf{C}}$), interaction intensity threshold ($T_{\mathbf{I}}$) and error threshold (T_{ϵ}). For any particular spot to be retained, the negative control value had to be less than $T_{\mathbf{C}}$ and either the intensity value had to be greater than $T_{\mathbf{I}}$ or the error value had to be less than T_{ϵ}. More formally these criteria are given by the formula:

$$(\mathbf{C} < T_{\mathbf{C}}) \wedge [(\mathbf{I}_i > T_{\mathbf{I}}) \vee (\hat{\epsilon}_i < T_{\epsilon})]$$

The criteria for determining which human data to retain were more complicated. Rather than only using a cutoff for the negative control values, a spot which has a negative control value above $T_{\mathbf{C}}$ could

still be retained if the interaction value was higher than the negative control value, within an acceptable margin given by the constant α where $\alpha > 1$. This gives the formula for deciding which human data to keep:

$$[(\mathbf{C}_i < T_{\mathbf{C}}) \vee (\mathbf{I}_i > \alpha \mathbf{C}_i)] \wedge [(\mathbf{I}_i > T_{\mathbf{I}}) \vee (\hat{\epsilon}_i < T_\epsilon)]$$

3.3.3.5 Partitioning data

Once data had been normalised and unreliable values removed, the data was partitioned into SIMs and non-SIMs by taking an interaction threshold and defining all peptides above this threshold as interactors. This threshold was set above a large cluster of peptide interaction values near 0. For the *Arabidopsis* data this value was 0.125 and for the human data 0.15.

3.3.4 SUMO site data

Training data to build the SUMO site predictors was gathered from Ren *et al.* (2009) and included the data used to train SUMOsp version 2. The authors provided Uniprot database identifiers and the indices of the SUMOylated lysines but not any sequence data. These identifiers were collected, however, a number of the database entries had been retired and replaced with more accurate sequences. These updated IDs and corresponding protein sequences were first checked with an R script to ensure all lysine indices given by Ren *et al.* (2009) still matched lysines in the new sequences. For a number of the new sequences, the given indices did not match lysines but were within a few amino acids of one. For each of these index mismatches, the index was corrected to the nearest lysine. As had been done by Ren *et al.* (2009) and Teng *et al.* (2012), all lysines not identified as SUMO sites were used as negative data for training predictors. Sequences were then collected which included 5 amino acids both downstream and upstream of the central lysine. Redundant sequences were then removed so that every training sequence was unique. In total 8318 sequences were collected, 332 SUMOylated sequences and 7986 non-SUMOylated sequences. It is likely that the non-SUMOylated set will contain some false negatives since a fraction of the non-surface exposed motifs would still be suitable substrates for SUMOylation. This issue however, has not had a significant effect in previous predictors and random forest models are resistant to incorrectly labeled training data.

3.3.5 Peptide and protein analysis

3.3.5.1 Amino acid conservation scoring

To calculate the similarity between a number of sequences, the sum of pairs method was used as this takes amino acid properties into account by giving a score to each possible substitution (Pei & Grishin, 2001). To score substitutions, the homologous structure derived substitution (HSDS) matrix from Prlić *et al.* (2000) was used. The HSDS matrix was derived from comparisons of domains with similar structure but with low sequence identity with intent to develop a scoring matrix that is applicable for scoring distantly related sequences. The rationale for using this matrix was that this work is more concerned with the structure and function of peptides than the evolutionary relationship between sequences. One modification was made to the matrix: since it was intended to be used for sequence alignment it did not penalise gaps in an alignment heavily, the score for pairs with a gap was reduced to 0. The penalty was changed as, for the purposes of scoring structural similarity, gaps do not represent conservation of amino acids.

To calculate the similarity of a set of n aligned sequences, which are represented by a two-dimensional matrix $\mathbf{A}$, the similarity, S, at the k^{th} aligned amino acid position is given by

$$S_k = \left(\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \mathbf{H}(\mathbf{A}_{k,i}, \mathbf{A}_{k,j}) \right) / \binom{n}{2}$$

where $\mathbf{H}$ is the substitution matrix.

The computational time for this algorithm grows at rate of $O(n^2)$ for the number of sequences and the algorithm becomes computationally intensive for multiple full length proteins. An R implementation of this calculation was too inefficient for the sequence predictor and was instead written in C and compiled as a shared library to be called in R (on Unix type systems only). The source code for the C and R functions are shown in Appendix C.2.1. Multiple sequence alignment was performed using the MUSCLE algorithm (Edgar, 2004).

3.3.5.2 Preference logos

Sequence logos are often used to analyse alignments of amino acids and nucleic acids. Sequence logos show which residues at which position are present at a rate above random chance, measured as a combination of Shannon entropy and residue frequency (Schneider & Stephens, 1990). While these graphs are useful for showing which residues are important, they do not show residues which are found at a lower than expected frequency at a position. To analyse SIM sequences, a more robust method was required that measured both over- and under-representation of amino acids at each position. Berry *et al.* (2006) developed a method to overcome the limitations of sequence logos by representing DNA letters on a graph whose height was the quotient of the observed and expected frequency.

The method from Berry *et al.* (2006) was taken as a starting template and adapted to make a method suitable for analysing SIMs, known as preference logos. One issue that this new method had to address was the low number of observations and large alphabet size, often leading to small numbers for counts of rare amino acids. To address this issue, the frequencies of the observed amino acids were subtracted from, rather than divided by the expected frequency. This reduced amplification of noise in the results, especially for under-represented amino acids. In addition, amino acids that did not show significant difference to the expected distribution were removed from the figure, leading to a figure that shows only the amino acids which deviate from the expected frequency. The Poisson exact test was used to determine which amino acids were not significantly different from the expected frequency and only amino acids with a $p < 0.05$ were kept. The Poisson exact exclusion method also had the benefit of removing data points with too few observations to show statistical significance. A multiple comparison correction for the p values was initially included using the Benjamini and Hochberg method but this was found to be too restrictive and was not used on the results presented in the chapter though ideally if the number of observations were higher, a multiple comparison correction should be used. The R source code for this function is shown in Appendix C.2.2.

3.3.6 Principal component analysis of amino acid features

To improve the performance of the SIM models, amino acid factors (i.e. the amino acid letters) were converted into numeric variables that describe the chemical properties of the different amino acids. These variables were generated from a principal component analysis (PCA) of amino acid chemical data. Amino acid chemical data were used from the AAindex database by Kawashima & Kanehisa (2000), accessed from the seqinR R package (Charif & Lobry, 2007), and PCA was performed using the FactoMineR R package (Husson *et al.*, 2013). The AAindex contained 544 different indices of properties for amino acids, 13 of these indices were excluded because they did not contain data for all 20 amino acids. The first 5 principal component dimensions were then saved for later use in machine learning models.

3.3.7 Random forest predictors

Training data collected for SIMs and SUMO sites were used to train random forest predictors. For each feature, data was divided into multiple subsets and a different random forest model was trained for each set. The R implementation of the random forest algorithm by Breiman (2001) was used (Liaw & Wiener, 2002).

3.3.7.1 Data treatment

The processed SIM data was split into six training sets, splitting by the two species (*Arabidopsis* & human) and by the three SIM types, A, B and R. The synthetic SIM sequences were sampled from a limited set of peptides with different amino acid distributions compared to natural peptide sequences. To prevent the models being biased by this, random sequences with natural amino acid frequencies were added to the training sets as negative data. For each SIM type, random sequences were generated by using natural amino frequencies as probabilities to randomly sample each amino acid in a 13-mer sequence. The same number of these randomly generated sequences was added to each SIM type set as there were non-interacting peptides. For the B type SIM random sequences, position 5 and 6 corresponding to the highly conserved DL motif was constrained to these amino acids as sequences tested with the random forest predictors for this motif would be prescreened to contain this conserved DL motif. If the amino acids in the random negative set were not constrained to DL at position 5 and 6, the random forest would select these amino acids as the most important variables, which would lead to poor performance as the important features outside the DL motif would not be prioritised.

For the SUMO site data, the positive training data was partitioned into type I and II sites, and all the negative data was added to each set. The SUMO site type was not known and was determined by performing k-means clustering on the positive data, with $k = 2$. To perform the clustering, a distance matrix was calculated by taking three amino acids downstream and upstream of the central lysine (K ± 3) and converting these to the numerical values from the first dimension of the PCA of amino acid features. The Euclidian distance was then calculated for each pair of vectors to generate the distance values. To determine the SUMO site type of the sets of data from the cluster analysis, the resulting groups were analysed for conformity to the canonical ΨKx[E/D] motif, with the closest to the group being designated as type I.

Next the datasets were prepared for the random forest models by converting the amino acid factors into numeric vectors from the PCA of amino acid indices. Each amino acid in the 13-mer SIM sequences was converted into 5 PCA dimensions resulting in a vector of 65 dimensions for each SIM sequence in all of the SIM subsets. For the SUMO site data, 11-mers consisting of five amino acids downstream and upstream of the central lysine (K ± 5) were converted into the 3 PCA dimensions, resulting in vectors of 33 dimensions for each SUMO site sequence in each subgroup.

3.3.7.2 Building and optimising random forest predictors

A multistep approach was taken to build the random forests for each data subset. Since each sub-dataset contained many times more negative training data than positive (for both the SIM and SUMO site data), subsampling of the data was required to prevent the random forest models optimising for the prediction of the negative data at the expense of positive data, i.e. the random forest models would tend to a

specificity of 100% and a sensitivity of 0% without subsampling. The internal random forest sampling method was used, maintaining the integrity of the OOB error estimation. An approximate sampling ratio was determined by exhaustive searching to find the ratio where the OOB error estimate of positives and negatives was the most similar, which was generally close to a ratio of 1:1.

Once the optimal subsampling ratios had been determined, a very large random forest with 10 000 trees was trained on the full datasets for each training subset with purpose of calculating variable importance. The resulting variable importance, measured in mean decrease in the Gini coefficient, was used to rank the importance of each variable in the input vector.

Next parameter selection was performed with another exhaustive search. A parameter search with two variables, v and m was performed with each data subset. The variable v is the number of highest ranked variables to be used and m is the number of variables used at each node in each tree. The range for m was 1 to 10, and for v 1 to the maximum number of variables in the training data, where $m \leq v$ in all cases. A small random forest of 250 trees was trained and the performance of the random forest was estimated by calculating the OOB estimate of the AUC using the R package ROCR (Sing *et al.*, 2005). The score parameter used in the AUC calculation was the proportion of trees predicting a positive value. This was then repeated 25 times and the mean AUC value with the 95% confidence interval for each combination m and v was calculated. Algorithm 2 details this process. 10 performance curves were generated for each training data subset, with a specific curve for each m value. The m and v pairs that generated the maximum mean AUC values were then taken as the optimal parameters for each data subset to use to train each random forest predictor.

Algorithm 2 Algorithm for finding optimal random forest parameters. RF is the random forest function; v is the number of variables with a maximum of N_v; m is the number of variables to use at each tree node with a maximum of N_m and $\mathbf{T}$ is a matrix of training data.

For each training data subset perform the following to calculate a vector of means and confidence intervals.

for v *in* 1 *to* N_v **do**
  for m *in* 1 *to* $min(v, N_m)$ **do**
    $\mu_{v,m} =$ mean AUC of $RF(\mathbf{T}, v, m)$
    $\mathrm{CI}_{v,m} = 95\%$ confidence interval of $RF(\mathbf{T}, v, m)$
  end
end

The final random forest predictors were then trained with the calculated optimal m value and the v most important variables, adding trees until the OOB estimate of error could not be improved any further, which in all cases was around 2000 trees. The resulting RF models, along with metadata about the sequence features, was encapsulated into a sequence feature object. The metadata included which

variables were used, the type of sequence feature, the indices of the core and a search mask. These metadata were used to correctly configure the predicator for each sequence feature. For SIMs the core corresponds to the central hydrophobic patch and in SUMO sites the central lysine. The mask is a short regular expression that determines which subsequences of a full length protein are tested with the sequence feature predictor. For the SUMO sites, the mask only allowed sequences with a lysine at position 6 within the subsequence, the same position as the SUMOylatable lysine in the training data. For the SIM prediction models the mask matches core features; the SIM type A and R mask matches three hydrophobic residues in the 4 residue cores (ΨΨxΨ or ΨxΨΨ) while for SIM type R the core matches the immutable DL amino acid pair. The masks were used to decide which subsequences to test within a protein sequence and this method dramatically reduces the number of subsequences tested, reducing the computational time required for the predictor to run. Sequences that do not contain the core features in these masks are very unlikely to be SIMs.

3.3.7.3 Quantifying and comparing random forest performance

Once optimal random forest model parameters had been found for each data subset, the performance of those models was assessed by calculating the OOB ROC statistics. RFs were trained 25 times and the mean OOB AUC and 95% confidence intervals were calculated. The SUMO site predictors were compared with the other published predictors, SUMOsp 2 and seeSUMO. The full training set used to build the random forests was queried against both predictors, using their web interfaces. For SUMOsp, the training data were split into type I and II sites, as the output of this model distinguishes between these two types of site. The score thresholds were set to their lowest values so that the predictors would return score values for all training data. The resulting score values and their corresponding interaction values were used to generate ROC curves and calculate the AUC for each predictor.

The score used from the random forest models is the proportion of trees giving a positive prediction for a given peptide sequence. These score values do not however provide any useful information about the accuracy of the prediction these models give, and using OOB ROC data the false positive rate (FPR) was modelled using an inverse sigmoid for each random forest model. The *FPR* function is given by

$$FPR(score) = \frac{1}{1 + e^{(\alpha \cdot score + \beta)}}$$

with the constants α and β estimated by performing Gauss-Newton non-linear least squares regression on the estimated ROC data. An inverse FPR model was used to calculate score cutoffs for the random forest models. This function is given by

$$score(FPR) = \frac{\ln(\frac{1}{FPR} - 1) - \beta}{\alpha} \qquad \begin{cases} \text{if } FPR > \frac{1}{e^{\beta}+1} & , FPR = \frac{1}{e^{\beta}+1} \\ \text{if } \frac{1}{e^{(\alpha+\beta)}+1} \leq FPR \leq \frac{1}{e^{\beta}+1} & , FPR = FPR \\ \text{if } FPR < \frac{1}{e^{(\alpha+\beta)}+1} & , FPR = \frac{1}{e^{(\alpha+\beta)}+1} \end{cases}$$

and uses the same α and β constant values calculated for the FPR model.

3.3.8 HyperSUMO, a graphical user interface sequence feature predictor

A web-based tool called HyperSUMO was developed to release the SUMO-related sequence feature (i.e. both SUMO sites and SIMs) predictor to the wider research community. The shiny application framework (RStudio, 2014) was used to develop the web tool, which was designed to run on a 64 bit linux server. HyperSUMO takes a set of FASTA formatted sequences and first aligns these, plots a graphical representation of the alignment and then over-lays predicted SUMO sites and SIMs onto the alignment. The graphical representation also contains a graph of sequence similarity along the alignment calculated. The results are also displayed in a table, with the location, sequence, type and confidence. The tool has the option to select one of three cutoff thresholds to control the FPR of the prediction as well as select between *Arabidopsis* and human SIM predictors.

To manage and align the protein sequence data, the bio3d R package (Grant *et al.*, 2006) was used. To read data from the text input box in the graphical user interface (GUI) a modified version of the bio3d FASTA format sequence reader was developed. The sequence reader first checks that the correct format is used and flags issues such as the use of ambiguous characters which were not supported or for presence of nucleic acid sequences. A sequence is determined to be a nucleic acid if it contains at least one of each nucleobase letter and no amino acid specific letters. If multiple sequences are uploaded they are aligned with the MUSCLE algorithm (Edgar, 2004) and a conservation graph is calculated using the sum of pairs method described earlier.

After query sequences are submitted, HyperSUMO uses the RF predictors to scan along each sequence in the input and generate sub-sequences that match the masks encapsulated within the RF models. The amino acids in these sequences are then converted into the numeric PCA dimensions and the required variables for the model are selected. Converted sub-sequences are then sent to the RF predictors which return prediction scores for all sub-sequences. Any subsequences with a score above the user selected threshold are then indicated on the graphical alignment and the sequence data is added to the output table. For the SUMO site predictors, if both the type I and II RFs predict a positive result at the same lysine, the type I site overrides the type II, which is discarded; this is done as the type II predictor can predict both types of site but has worse accuracy than the type I site predictor. For each positive

result, an approximate confidence value is calculated corresponding to the specificity. Using the sigmoid FPR models, the specificity is calculated as $Sp = 1 - FPR(score)$. Source code for HyperSUMO will be provided on request.

3.3.9 Genome-wide screen for SIMs

The random forest SIM predictors were used to perform a genome-wide screen for likely SIM containing proteins in *Arabidopsis*, integrating multiple sources of information to improve the accuracy of the predictions. The SIM screen used evolutionary information to find candidate SIMs that were conserved and did not lie within any predicted functional protein domains. Domain sequences were removed as these tend to be structured while SIM sequences most often occur within unstructured regions of proteins.

To incorporate the evolutionary information into the screen, predicted *Arabidopsis* orthologous genes were used from the work by Baxter *et al.* (2012) which defined reciprocal BLAST hit gene pairs as orthologs. Orthologs from the following 7 species were used: *Populus trichocarpa* (poplar tree), *Medicago truncatula*, *Vitis vinifera* (grape), *Phoenix dactylifera* (date palm), *Musa acuminata* (banana), *Sorghum bicolor* and *Oryza sativa* (rice). From those data, *Arabidopsis* proteins with at least 2 orthologs with non-ambiguous sequences were grouped and then aligned using the MUSCLE multiple sequence alignment algorithm (Edgar, 2004). Each set of aligned proteins, corresponding to an *Arabidopsis* gene, was then analysed. Each sequence in the alignment was then scanned with each SIM predictor, resulting in a SIM score matrix for each SIM type. For each SIM score matrix, the values were converted into FPRs using the sigmoid model discussed earlier. The mean of the FPR was then calculated at each amino acid position. The score values of gaps in the aligned sequence were set to 0, or an FPR of 1.0, giving a very high penalty to these sequences. Once the mean SIM FPR scores and conservation scores at each position in the alignment were calculated, values that corresponded to alignment gaps in the *Arabidopsis* sequence were removed from the dataset as they were not meaningful in the context of screening the *Arabidopsis* genome.

Next, amino acids that were within predicted protein domains were identified. The entire *Arabidopsis* translated genome was uploaded to the NCBI conserved domains database batch processor with a cutoff of $p < 0.01$ (Marchler-Bauer *et al.*, 2013). Any amino acids in *Arabidopsis* proteins that were predicted to be within a domain superfamily were flagged. Then for each protein, the number of predicted SIMs outside a domain region and with mean FPR ≤ 0.2 were counted. From these data, 500 proteins with the lowest FPR SIM scores were used, with the FPR cutoff of 0.1021. The top 500 predicted SIM-containing genes were then analysed for gene ontology (GO) term enrichment using the Virtual Plant internet interface (Katari *et al.*, 2010). Biological function and molecular process ontologies were analysed. To generate an overview of the results from the GO term analysis, the terminal node keywords

in the GO term hierarchy were used with a cutoff of $p < 0.001$.

3.4 Results

3.4.1 SIM peptide array image analysis

To generate the SIM data, three separate SIM arrays were manufactured with a total of 1200 peptide spots. The first two arrays (SIM array parts I and II) contained 600 non-modified peptide sequences for all three SIM types while a third (phosphorylated SIM array) contained 600 peptides with various combinations of phosphorylated serine, threonine and tyrosine residues. The unmodified and phosphorylated arrays were manufactured on different runs on the peptide array synthesiser and staining with Ponceau-S (Figure 3.2) showed that there were fewer successfully manufactured peptide spots on the first two arrays (SIM array parts I and II) compared to the phosphorylated peptide array. This difference in efficiency was most likely due to differences in manufacturing conditions rather than differences in the peptide sequence as the same templates for generating peptides were used to design the peptides. Due to the lower peptide manufacture efficiency in the first two arrays, many more peptide spots were excluded from the processed dataset than from the phosphorylated array.

The intensities of the Ponceau-S stained peptide spots were used to normalise the intensities of the spots from the SUMO far-western blots. The associated error for each correction was calculated based on the intensity of the Ponceau-S peptide spot and is shown in Figure 3.3. This correction error was used as the basis to decide which peptide spots to exclude from the data to ensure that false negatives were not included in the final dataset. Without this error correction, peptides that would have interacted with SUMO but that were not synthesised in sufficient amounts would have shown no signal in the far-western blots and would have been labeled as non-interacting peptides. Having such false negatives in the predictor training data would have resulted in reduced performance.

The images of the SUMO far-western blots had uneven backgrounds in both the x and y directions of the images and without correction, measured intensity values would have been spatially biased. The method to flatten the base line successfully corrected the spatial background defects and set the baseline values of the images to 0. Figures 3.4 to 3.7 show the results of the baseline flattening from samples taken in the x and y directions of all far-western blots. The areas where these samples were taken from are shown in the Appendix figures B.1 to B.4. The data for the HsSUM1 far-western blots (Figures 3.6 & 3.7) was noisier and the baseline flattening was less effective than that of the AtSUM1 far-western blots (Figures 3.4 & 3.5). Some regions of the baseline under- or over-shot the 0 line however, the magnitude of these errors was small enough not to have a significant effect on the intensity values.

The normalised far-western blot images of the AtSUM1 arrays are shown in Figure 3.8. Of the two SUMO isoforms screened, these arrays produced less noisy data as there was very little signal outside

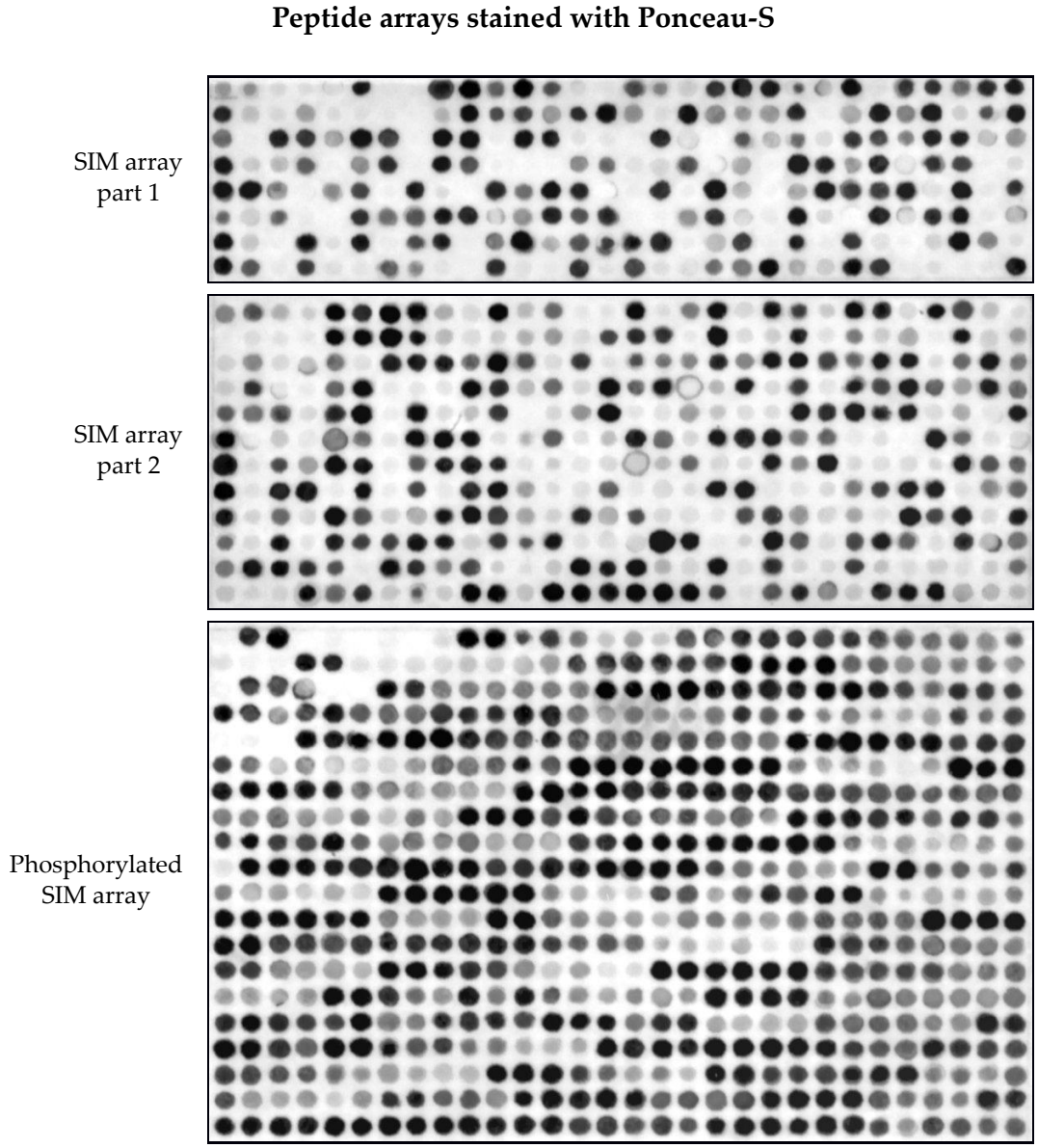


Figure 3.2: Ponceau-S stained peptide arrays. Staining shows the approximate amount of peptide at each spot. Staining shows that a large number of peptide spots were not synthesised effectively and the original SIM array (in two parts) had a worse efficiency than the later phosphorylation array. The intensities of the peptide spots were used to normalise far-western blot results. Data for spots with little or no peptide were excluded from the dataset.

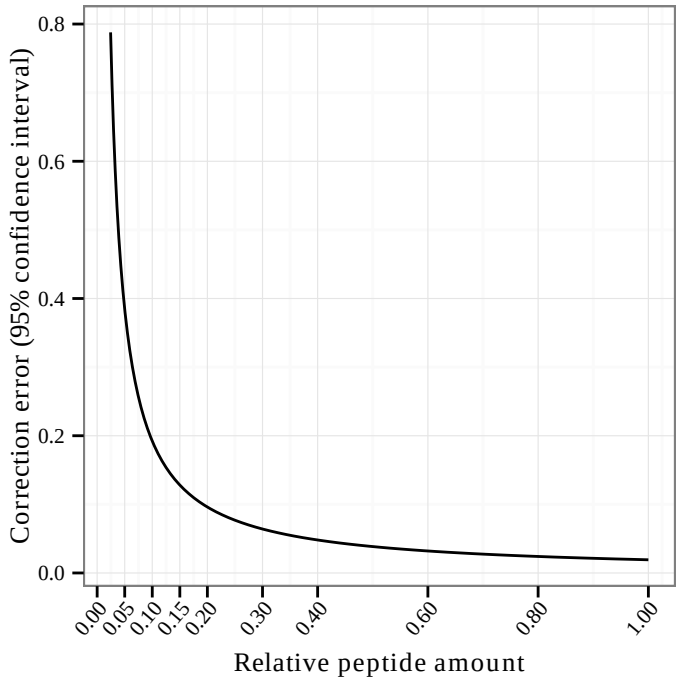


Figure 3.3: Correction error. Shows the 95% confidence interval of the error for the final interaction value after normalisation with peptide amount.

of the peptide spots and there were very few non-specific interactions in the control images. The control array was exposed to X-ray film for twice as long to ensure that any non-specific interactions were detected. Five non-specific were interactions identified in total, along with one spot that was determined to be an artefact (control SIM array part II).

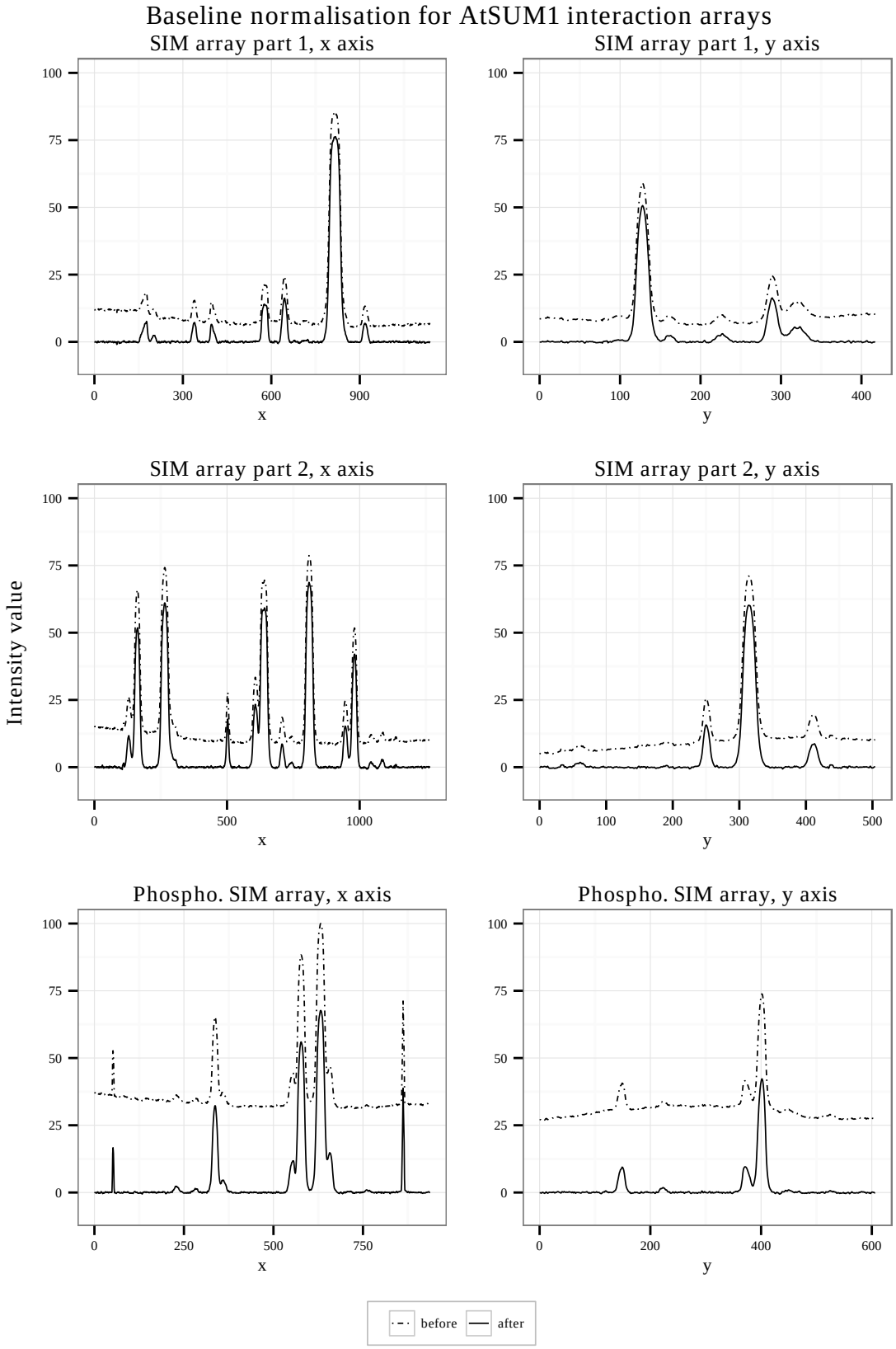


Figure 3.4: Results of baseline normalisation for AtSUM1 interaction array images. Normalisation flattened undulations and centred the baseline on zero. A sample of the baseline was taken in the *x* and *y* direction with the width of 1 peptide spot. The peaks in the data are caused by interacting spots in the array.

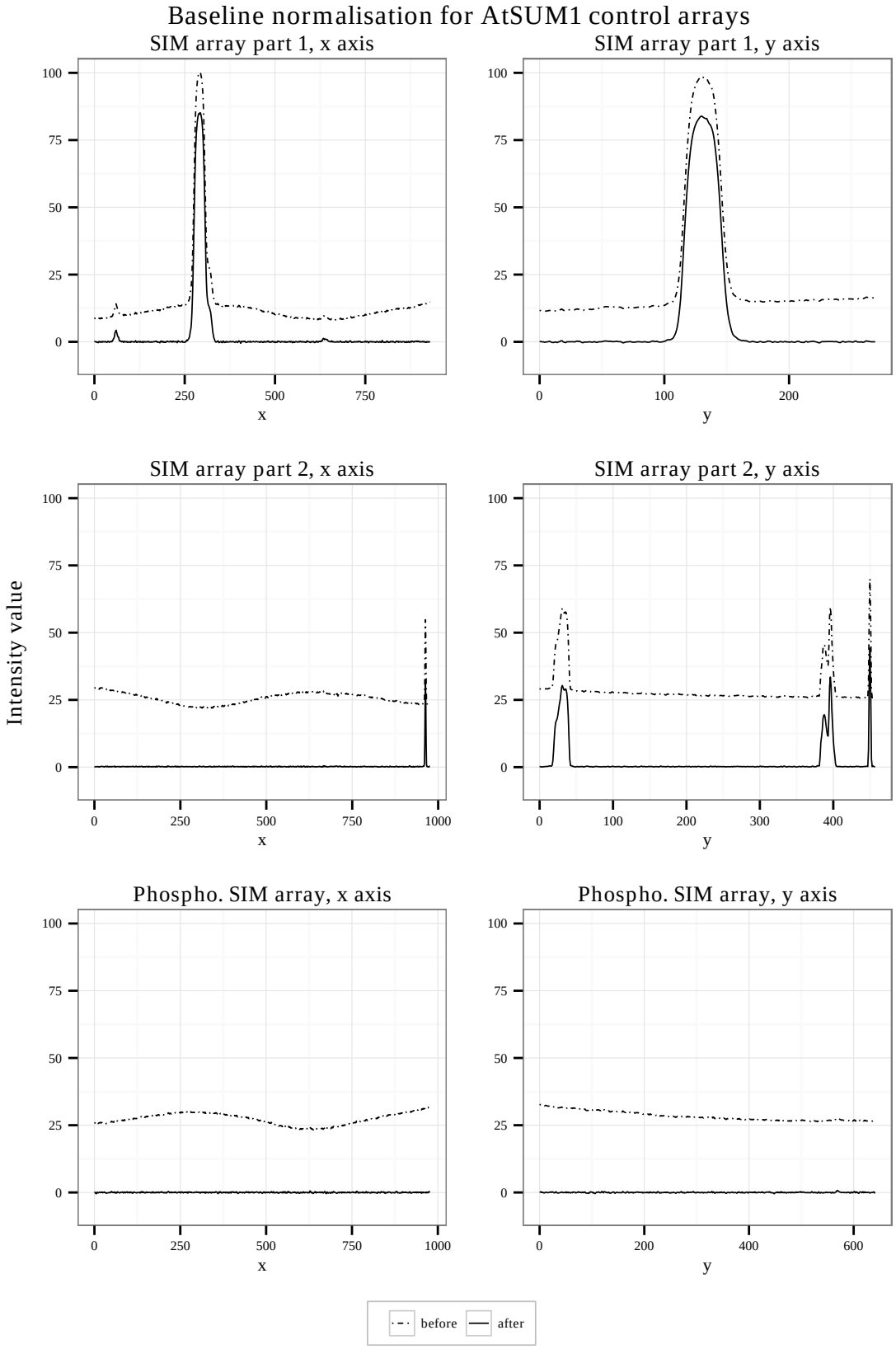


Figure 3.5: Results of baseline normalisation for AtSUM1 control array images. Normalisation flattened undulations and centred the baseline on zero. A sample of the baseline was taken in the *x* and *y* direction with the width of 1 peptide spot. The peaks in the data are caused by interacting spots in the array.

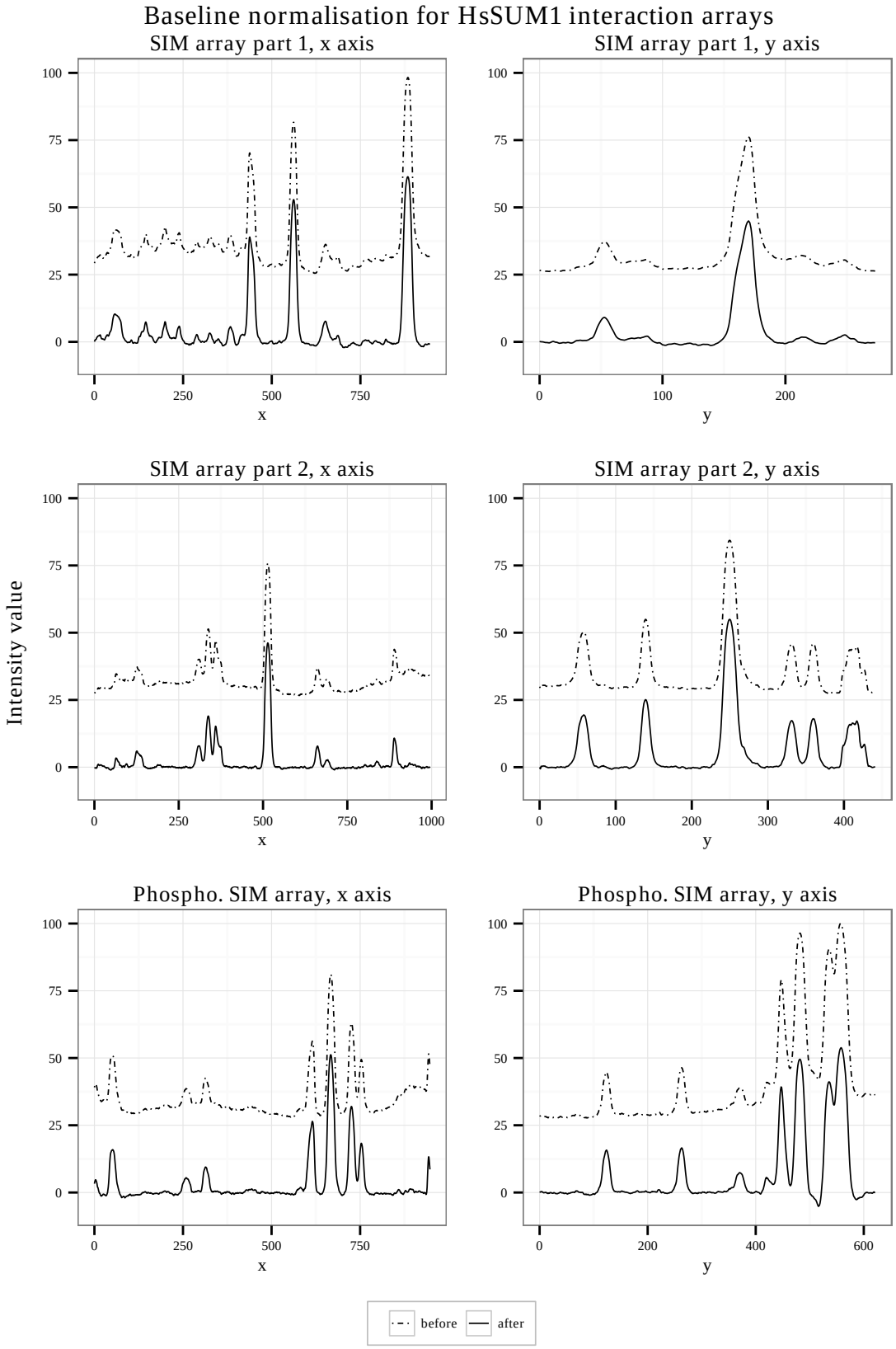


Figure 3.6: Results of baseline normalisation for HsSUM1 interaction array images. Normalisation flattened undulations and centred the baseline on zero. A sample of the baseline was taken in the *x* and *y* direction with the width of 1 peptide spot. The peaks in the data are caused by interacting spots in the array.

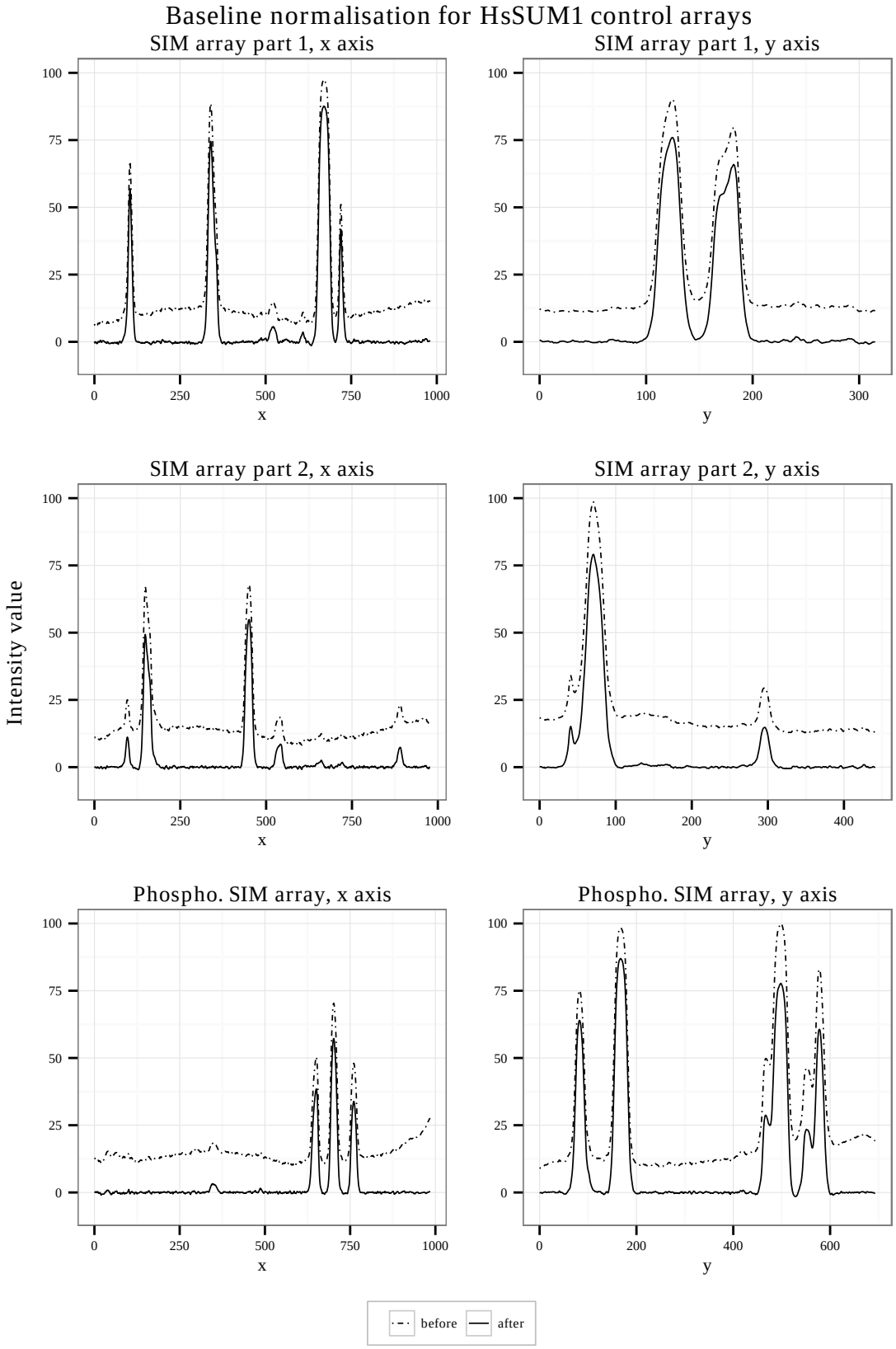


Figure 3.7: Results of baseline normalisation for HsSUM1 control array images. Normalisation flattened undulations and centred the baseline on zero. A sample of the baseline was taken in the *x* and *y* direction with the width of 1 peptide spot. The peaks in the data are caused by interacting spots in the array.

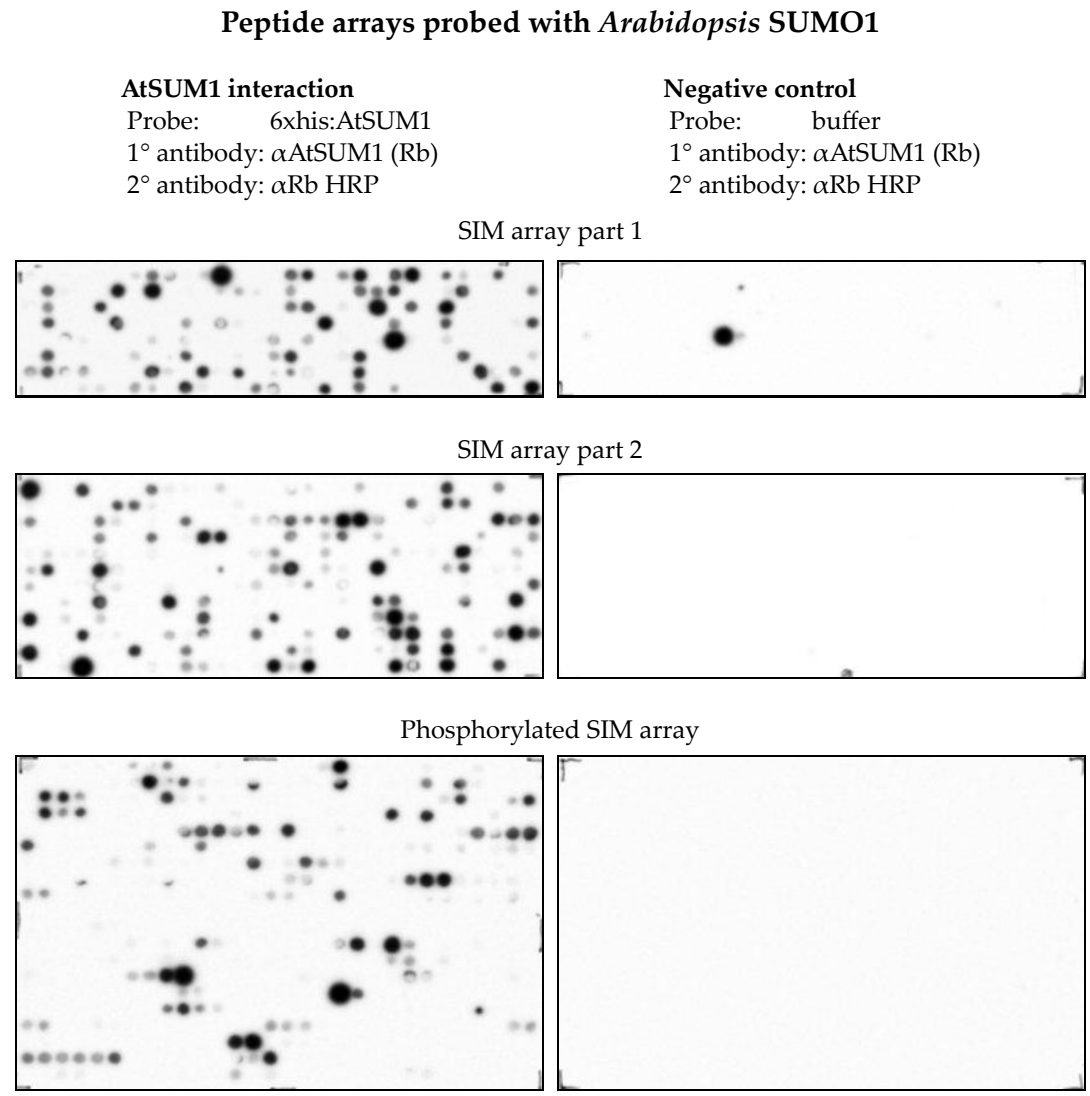


Figure 3.8: Far-western blot of peptide arrays probed with AtSUM1. For the negative control, the arrays were probed with buffer without AtSUM1 protein. Dark spots indicate an interaction. Only a few peptides showed interaction in the negative control blot and were removed from further analyses. The negative control blot was also exposed for twice as long to ensure that false positives were amplified. The dark spot in the negative control array actually corresponds to a very weak spot in the interaction array; this could be explained by less competition by the peptide for antibodies than the interaction array. This shows that the negative control blot was very sensitive to false positives providing a high level of confidence that the detected interactions were genuine.

Of the 5 non-specific spots, only one had a strong signal (control SIM array part I) and the intensity of this signal was many times that of the corresponding spot in the interaction far-western blot, showing that the controls were highly sensitive to non-specific interactions and giving a high confidence that any non-specific interactions were detected. The purpose of the control far-western blots was to exclude false positives from the SIM interaction datasets.

The HsSUM1 far-western blot images (Figure 3.9) were much noisier as there was signal in the inter-peptide spot space and a very large number of peptides had signal in the negative control blots. The most likely reason for the poor quality of the data was the primary antibody used in the assay,

αGST, which had a high level of non-specific binding.

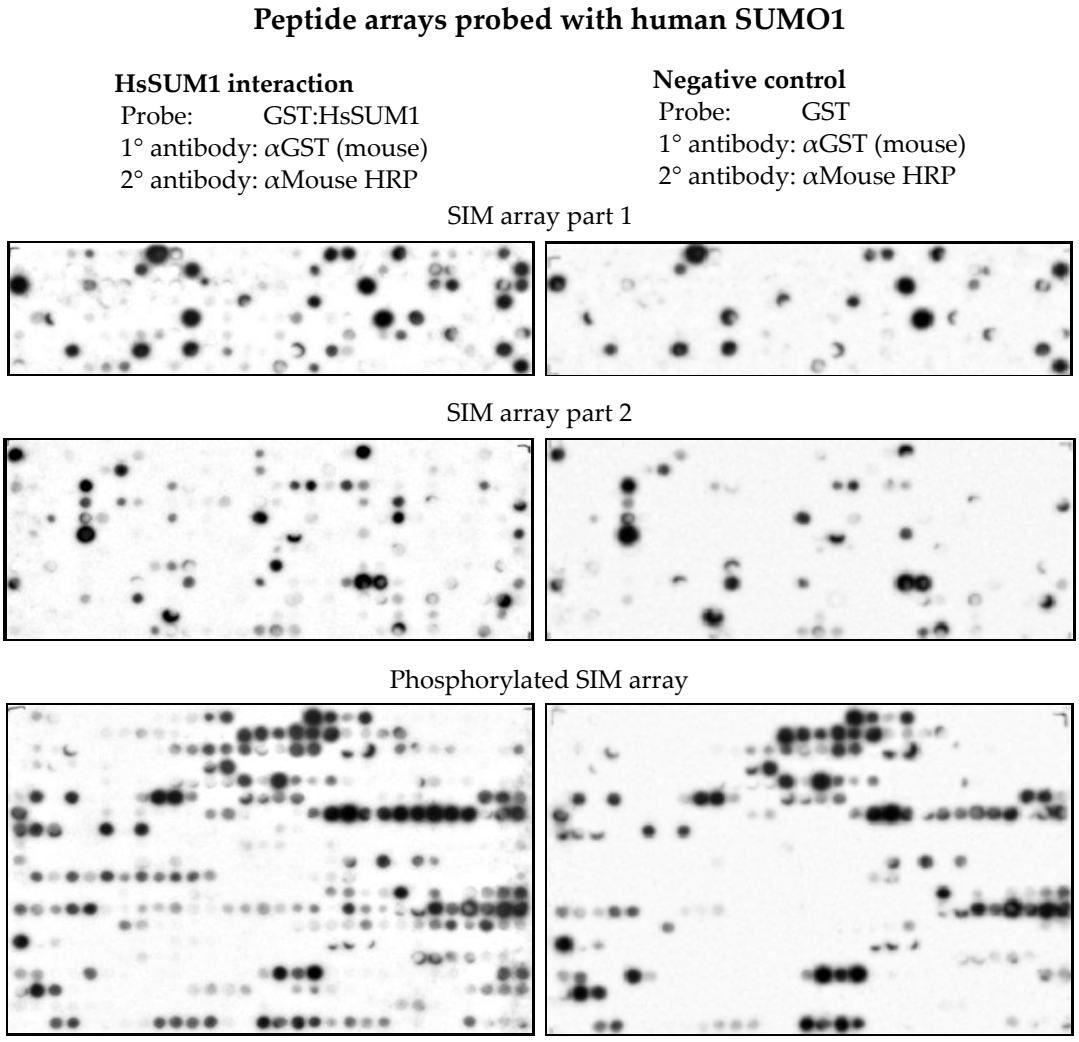


Figure 3.9: Far-western blot of peptide arrays probed with HsSUM1. For the negative control, the arrays were probed with GST protein, as the HsSUM1 used in the interaction array was fused to GST. Both array images were exposed to X-ray film for 5 minutes. There were a high number of non-specific interactions in these images.

Initially αHsSUM1 antibodies were used but these produced such a high signal in the control blots that no genuine interacting peptides could be identified (data not shown). αGST antibodies performed better, allowing identification of some genuine interacting peptides but the performance of the antibody was still far from ideal as about a third of the peptides had to be excluded.

The final results of the data processing and peptide exclusion are shown in Figure 3.10. Data-points that were excluded from the final dataset due to either too low peptide amount or signal in the negative control blots are shown in red. Overall a significant proportion of the peptides in the HsSUM1 far-western blots were excluded and, for both SUMO isoforms, more data was excluded in the unmodified peptide arrays (see Table 3.2). For the full list of peptides after processing see Table B.1 in the Appendices.

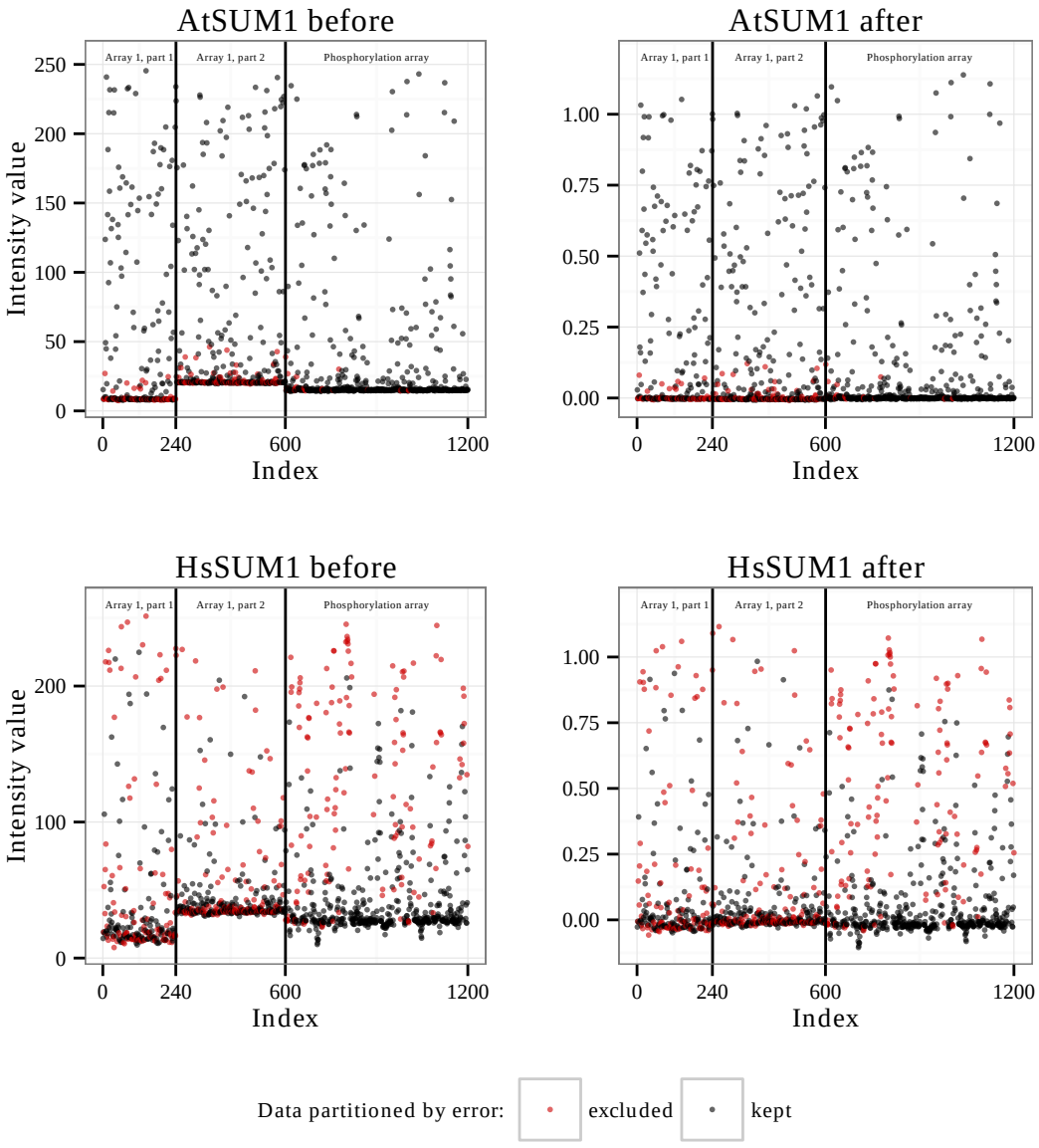


Figure 3.10: Processing results showing before and after values of the spots. Spots above the error thresholds that were removed from the dataset are shown in red, as well as exclusions by negative control. Each graph is partitioned into the three separate array pieces. Before processing the baselines of the different array sections were not aligned, after processing they are aligned and centred on zero.

	Number of peptide spots	AtSUM1 data excluded (%)	HsSUM1 data excluded (%)
SIM array part I	240	27.9	42.9
SIM array part II	360	25.6	38.6
Phosphorylated SIM array	600	3.8	23.5
Total	1200	15.2	31.9

Table 3.2: Percentage of peptide spots excluded from final SIM datasets.

3.4.2 Sequence analysis

In total there were 598 non-phosphorylated peptides (both interacting and non-interacting) for the AtSUM1 set and 484 for the HsSUM1 set. Between these two sets, there were 472 common peptides and the number that bound to the two SUMO isoforms is shown in Figure 3.11. Remarkably, only a small number of peptides, 16, bound to both AtSUM1 and HsSUM1, with the majority of the interacting peptides binding to only one SUMO isoform. More than twice as many peptides bound to AtSUM1, suggesting that the *Arabidopsis* isoform has less stringent requirements for the peptides it binds to. These data suggest that human and plant SUMO proteins may have diverged in their function and they now have different binding preferences. This emphasises the need to develop specific predictors for different SUMO homologs between species and within species too.

To analyse the sequences of the interacting peptides, the preference logo method was used that shows which amino acids are both over- and under-represented in a peptide sequence. This method was used to highlight which amino acids cannot be tolerated at positions along the SIM sequences. To analyse the SIM sequences, two reference background amino acid frequencies were used, the amino acid frequencies of non-interacting peptides and the natural frequencies of the amino acids. Using the frequency of the non-interacting peptides as a background identified important amino acid differences in the synthetic peptide dataset that were responsible for an interaction with the two SUMO isoforms. On the other hand using the natural amino acid frequencies as a background shows the actual structure of the SIM motifs.

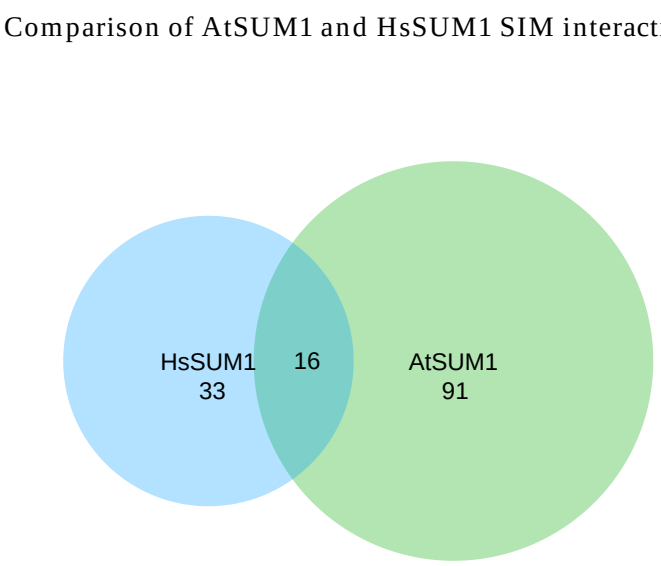


Figure 3.11: Venn diagram of AtSUM1 and HsSUM1 interacting SIM peptides. The data used to generate this figure was from a subset of peptides common to the AtSUM1 and HsSUM1 datasets. Very little overlap was seen in the sequences that the SUMO isoforms bind to, while AtSUM1 binds to about twice as many sequences in the dataset.

Due to the large size of the amino acid alphabet, a large number of sequences should be analysed to identify significant trends with the sequence data. For the AtSUM1 interaction set, there were enough positive sequences to clearly identify the major amino acids responsible for an interaction with AtSUM1 (Figure 3.12) however, due to the low number of positives for the HsSUM1 dataset, fewer amino acid features could be identified for these sequences (Figure 3.13). Overall the preference logos for AtSUM1 and HsSUM1 SIMs are very similar to SIM sequence logos published by Vogt & Hofmann (2012). However, because the set of peptides in the arrays was already similar to the published SIMs, this result should be considered cautiously and few if any conclusions can be drawn from it except that HsSUM1 results were more similar to the published SIM motifs than the AtSUM1 data. What is of more interest are the differences between the AtSUM1 and HsSUM1 interacting peptides.

The different SIM binding preferences between the two different SUMO isoforms lie outside of the hydrophobic core of these peptides. For SIMs A and B the core is from position 3 to 6 and for SIM R it is from 8 to 11 in the array peptides. The polar or charged amino acids aspartic acid, glutamic acid and serine are the most common amino acids upstream or downstream of SIMs A and R respectively, and this is seen generally in the peptide array data for both SUMO isoforms. The AtSUM1 sequences however have one very notable difference: at the position immediately next to the hydrophobic core (position 7 for both SIM A and R) the charged amino acids aspartic acid and glutamic acid are not tolerated, and the polar amino acid serine is much more common, which is in contrast to the HsSUM1 data and to published SIM motifs.

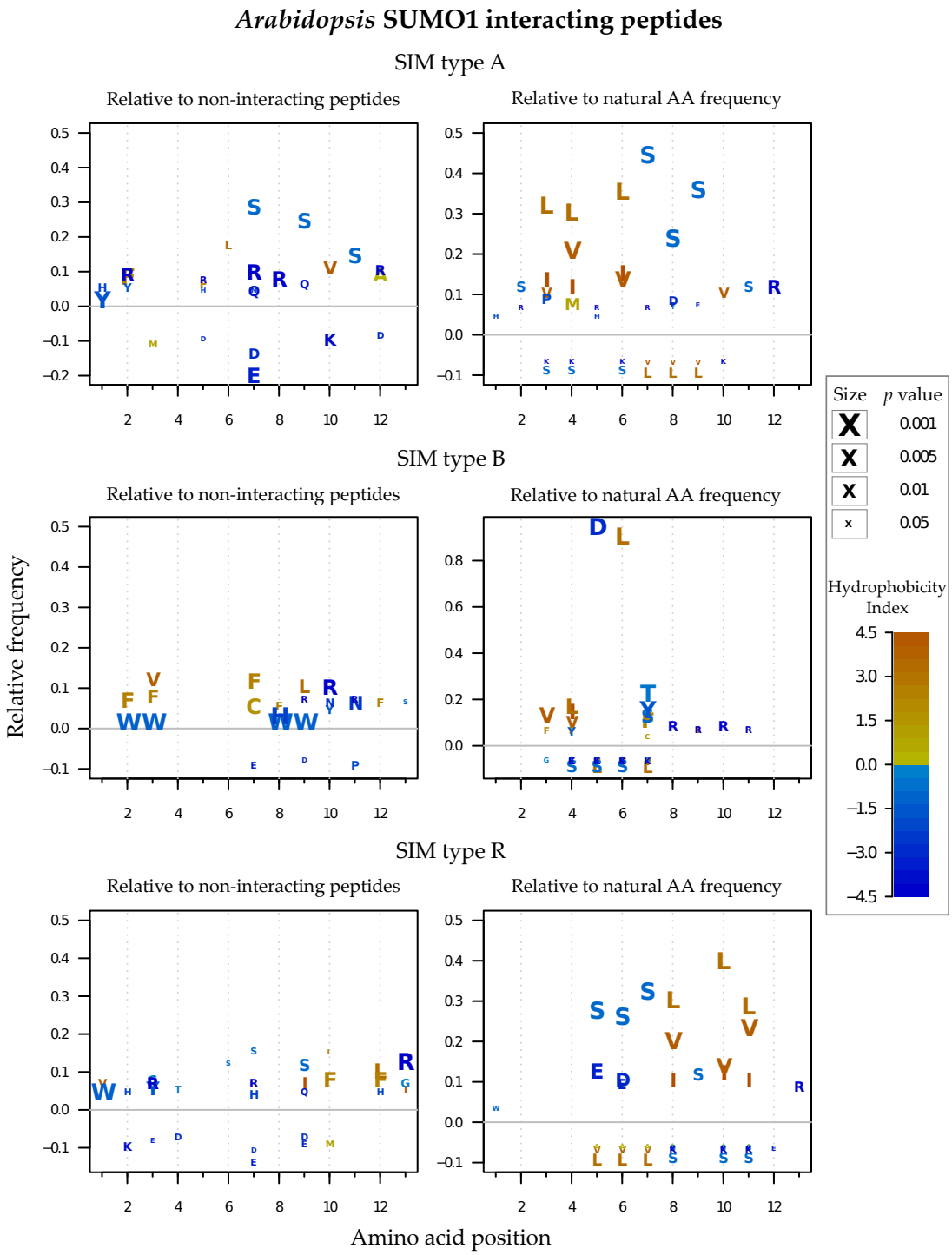


Figure 3.12: Sequence analysis of *Arabidopsis* SUMO1 interacting peptides.

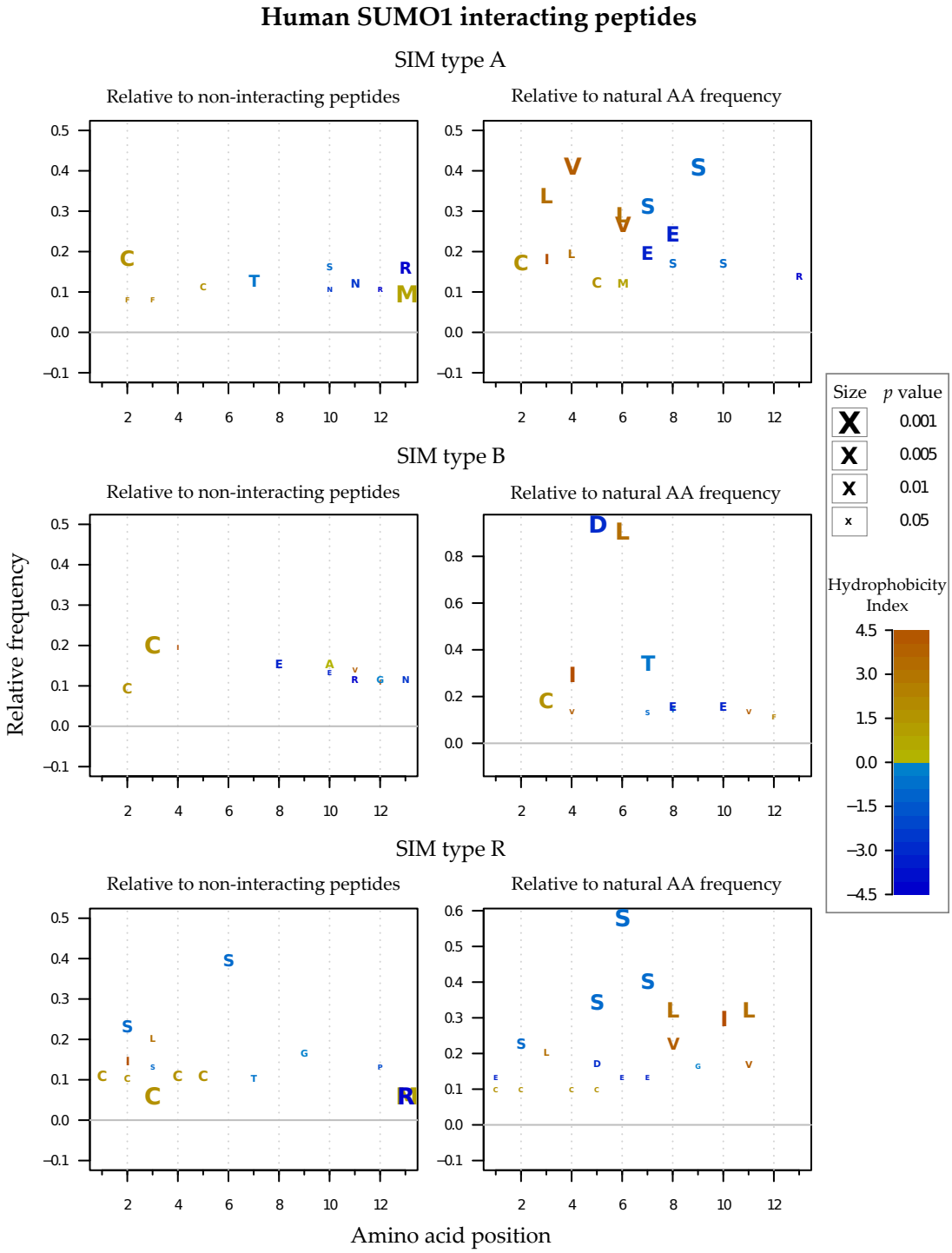


Figure 3.13: Sequence analysis of human SUMO1 interacting peptides.

The type B SIM again was very similar to published results for the HsSUM1 data but the AtSUM1 data showed some notable differences. The highly conserved polar threonine residue (SIM position 7), which forms part of the DLT motif, showed much greater variability in the AtSUM1 interacting peptides. This position appeared to be strictly polar for HsSUM1 and only allowed a threonine or serine residue, while for AtSUM1 interacting peptides also allowed a tyrosine substitution and substitutions with the more hydrophobic residues phenylalanine and cysteine. Upstream of the conserved threonine at position 7, the AtSUM1 interacting peptides also showed a preference for charged lysine residues at the next four positions, which was not the case for the HsSUM1 interacting peptides which instead appeared to favour glutamic acid, though the data for these peptides is limited due to the smaller dataset size.

Apart from the specific amino acid differences, there were also some notable differences in the frequencies of the amino acids in the set of interacting peptides between AtSUM1 and HsSUM1. In the HsSUM1 interacting peptide set, the frequency for cysteine was 4.1 times higher ($p = 0.063$) and aspartic acid was 1.3 times higher ($p = 0.01$), and in the preference logos for HsSUM1 interacting peptides a large number of cysteine residues are present in all motifs, especially toward the amino terminal (left hand side) of the motifs.

3.4.3 Phosphorylated SIM peptides

The effect of phosphorylation of the amino acids serine, threonine and tyrosine was investigated by designing a number of template peptides and randomly generating phosphorylated versions the base peptides. There are four possible outcomes to phosphorylation: activation where a noninteracting template interacts in the phosphorylated form, sustained interaction where both the template and phosphorylated peptide interact, deactivation where phosphorylation abolishes an interaction, and complete non-interacting peptides which do not interact at all. The results for the two SUMO isoforms are shown in Table 3.3. The majority of peptides screened did not interact with SUMO at all which is in concordance with previous results where only 10 - 20% of synthetic peptides showed interaction.

The effect of phosphorylation on peptide interaction was significantly different between the two SUMO isoforms, AtSUM1 and HsSUM1 (chi squared test, $p = 1.13E\text{-}10$), with phosphorylation having stronger interaction promoting effect with HsSUM1. A significantly larger number of AtSUM1 interacting peptides were deactivated by phosphorylation. Figure 3.14 shows a summary of the phosphorylated amino acid positions in SUMO interacting phosphorylated peptides. An interesting observation with AtSUM1 SIM A peptides was that phosphorylation of position 7 (most often a serine) in the peptides always resulted in loss of interaction, and this agrees well with the amino acid preference data which indicated that the negatively charged amino acids glutamic acid and aspartic acid were strongly selected against at this position as well. Phosphorylation imparts a strong -2 negative charge to modified amino

Type	Unmodified peptide interacts?	Phosphorylated peptide interacts?	AtSUM1	HsSUM1
Activation	No	Yes	21	28
Sustained	Yes	Yes	28	18
Deactivation	Yes	No	65	7
No interaction	No	No	304	257
Total			418	310

Table 3.3: Effect of phosphorylation on SIM peptide interaction. Table shows the number of times phosphorylation of test peptides had one of four effects. For AtSUM1, phosphorylation of peptides abolished interaction to a much greater extent than for HsSUM.

acids and these data suggest that position 7 of SIM A cannot tolerate a negative charge, while HsSUM1 can, though interestingly the position (8) in AtSUM SIM A can be phosphorylated without loss of interaction and can act as an activator. This trend however, is not seen in the reverse AtSUM1 SIM R at the corresponding amino acid position 7, where the negatively charged amino acids or phosphorylated amino acids were tolerated.

Due to the low number of individual phosphorylations screened for each amino acid in the peptide sequences, no statistically significant trends could be observed between the amino acid positions. This is not to say that no trends exist but that rather it is a limitation of having only a few positive observations for each position. However, what can be said is that phosphorylation of amino acids in the SIM peptides both retains an interaction and can act as molecular switch to regulate interaction. It is also notable that phosphorylation of amino acids either side of the hydrophobic core can maintain or activate interactions and in some cases the variable position, `x`, in the cores of SIM A and R (e.g. `LLxL`, `LxLL`) can also be phosphorylated, which was not initially expected since all phosphorylation switches currently published are due to phosphorylation of serines upstream of the hydrophobic core (Chang *et al.*, 2011; Hecker *et al.*, 2006; Percherancier *et al.*, 2009; Stehmeier & Muller, 2009). Overall the results suggest that there is greater diversity in the position of SIM phosphorylation than has been previously discovered.

3.4.4 Principal component analysis of amino acid indices

To build the random forest predictor models, amino acid factors were converted into numeric vectors of variables derived from a PCA of data from the AAindex database. This was to overcome the issue of correlation between variables in the index and to reduce the number of variables used in the random forest models. First the correlation of the AAindex was investigated. Figure 3.15, shows the correlation of a random sample of 50 variables out the total of 531, and quite a large number of these are highly correlated and so contain redundant information.

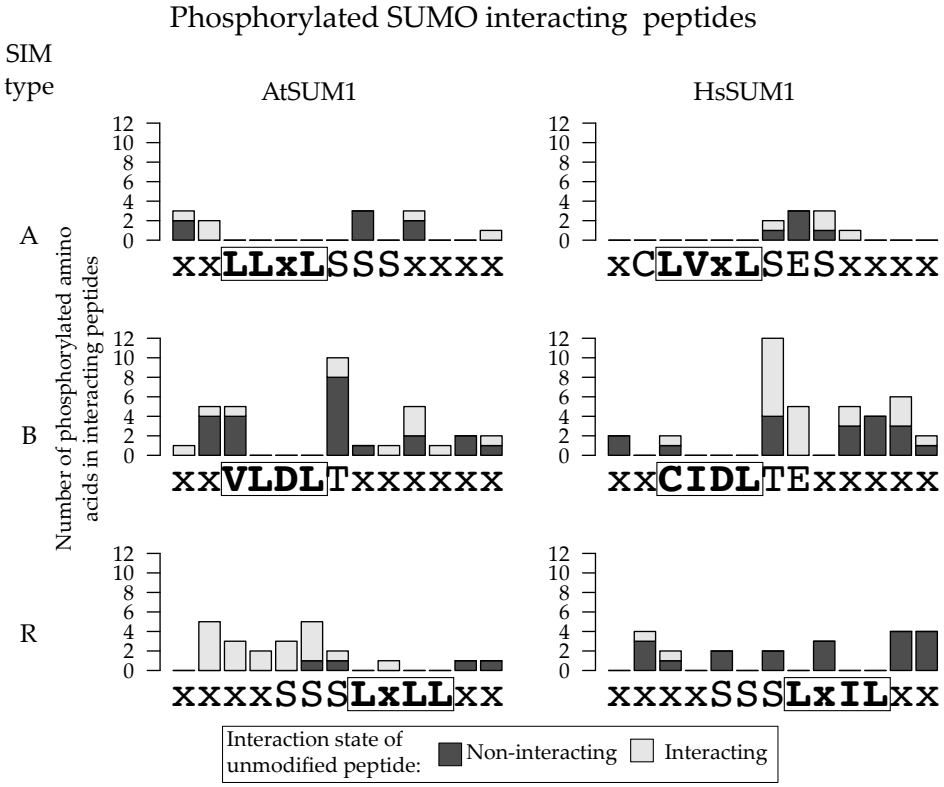


Figure 3.14: Summary of phosphorylated SIMs. Bars show counts of phosphorylated amino acids in SUMO interacting peptides with the consensus SIM sequence below; the boxed bold letters indicate the hydrophobic core. The amino acids specified in the consensus were not necessarily the amino acids phosphorylated as most positions are variable. The colour of the bars indicate whether the unmodified peptide interacted with SUMO: dark grey = no, light grey = yes. Dark grey bars indicate peptides that were activated by phosphorylation.

By performing a PCA on the AAindex variables, the variance between the 531 variables was accounted for in 13 principal components, with the first 5 of these accounting for the majority of the variance (Figure 3.16). To get a biological perspective on these principal components, the most similar variables in the AAindex to these components were found. Component 1 correlates well with hydrophobicity indices, component 2 with molecular weight, component 3 with hydrophobicity indices again but using different methods to determine the values, component 4 with partial specific volume and component 5 with steric restriction parameters.

3.4.5 SIM random forest models

The first step in building random forest models for the SIM data was to assess variable importance of the 65 input variables, which was generated by the algorithm, see Figure 3.17. There were very few trends or notable features in the importance of the variables apart from position 7 for *Arabidopsis* SIM type A at which aspartic acid and glutamic acid are not tolerated as they are in other species.

The difference in variable importance between AtSUM1 and HsSUM1 is probably due to the differences in size of the training sets and SUMO isoform binding preferences. The HsSUM1 models were

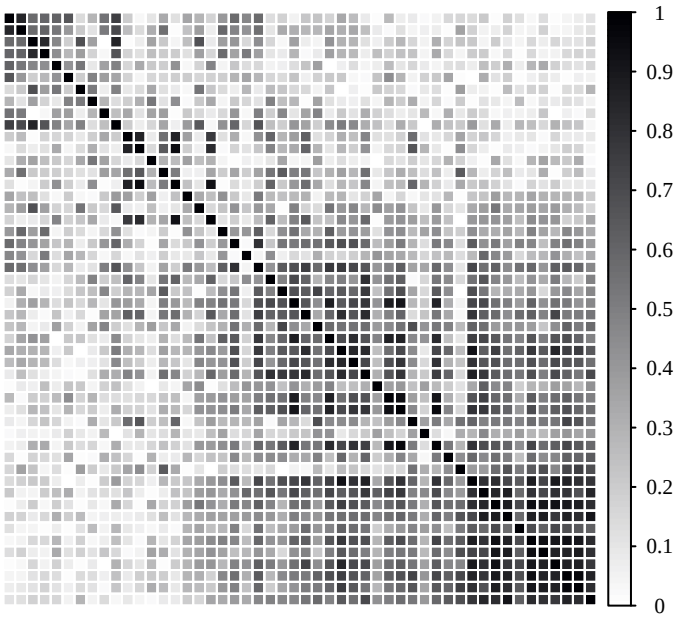


Figure 3.15: Correlation matrix of a sample of 50 amino acid indices from the AAindex database. Many of the amino acid indices are strongly correlated and so contain redundant information. The correlation values are on an absolute scale, showing either positive or negative correlation. The data in this figure was clustered to highlight the grouping of the amino acid indices.

more likely to be influenced by amino acid variance as the positive dataset was about half the size as that of AtSUM1.

Once variable importance had been determined, an optimisation algorithm was used to identify the optimal number of variables to use by assessing the performance of random forests adding one variable at a time starting with the most important. Also the number of variables used at each tree node, m, was optimised. The results of these optimisations are shown in Figures 3.18, 3.19 and 3.20 for SIM types A, B and R respectively.

For all of the random forest models, there was an optimal number of variables which resulted in a maximal performance measured as AUC; adding more variables after this point resulted in a decline in performance in all cases. The decline in performance with the HsSUM1 SIM models was more rapid than the AtSUM1 SIM models which is in concordance with the smaller dataset sizes for HsSUM1 positive sequences as variable scarcity is more problematic for these sets. The AtSUM1 SIM models also converged onto performance maxima with more variables than the HsSUM1 models suggesting that training datasets with more positives allowed more complex models to be supported than for the HsSUM1 datasets.

The optimal performance was achieved using a value of 1 for m for all models except the *Arabidopsis* SIM B model for which this parameter was 3 but there was very little difference in the performance of

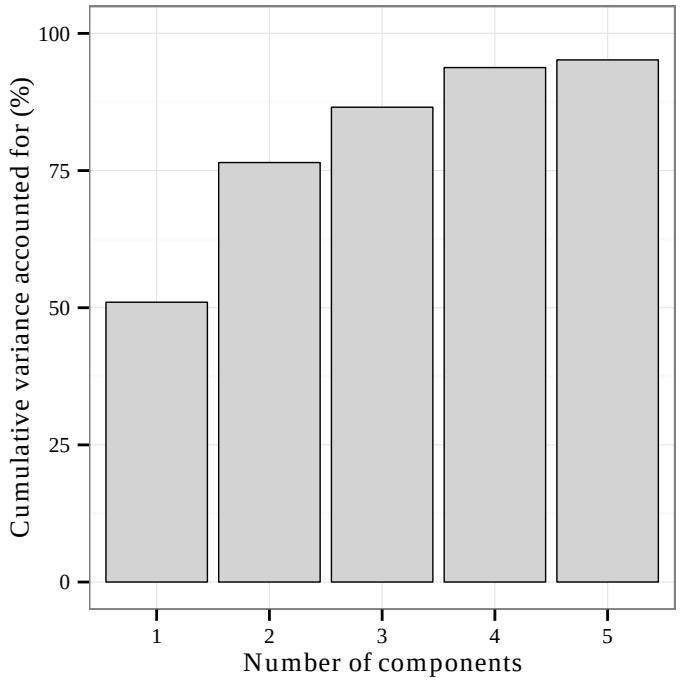


Figure 3.16: Cumulative variance accounted for in the first 5 principal components of the amino acid indices.

models with different *m* values for this SIM. These data show that in general for the best performance, sampling a single amino acid variable at each tree node results in better performance than sampling multiple amino acids.

Once the optimal variables and model parameter *m* had been identified, random forest models were constructed, adding trees until the OOB estimate of error could not be improved any further. For models the error stabilised at around 2000 trees and this value was used for all SIM types and species. The performance was assessed by building 25 instances of each model and calculating the AUC value for each model. These values are shown in Table 3.4. The variance for the AUC values was very small, showing that performance was very stable between different builds of the random forest models. For SIM A, the *Arabidopsis* version performed better while the human versions for the other two SIM types performed better, which is surprising given the smaller training dataset sizes for the human models. It may be the case that the calculated ROC AUC or any other performance measures are biased by small dataset sizes; while the OOB performance may be true for the training data used, it probably does not accurately reflect performance of the models when used on actual biological data. The ROC plots for the *Arabidopsis* and human data are shown in Figure 3.21 and show very jagged curves for the human models due to the small size of the training datasets. The ROC curves for the *Arabidopsis* models are more granular due to the larger size of the training datasets and should better reflect the true performance of these models.

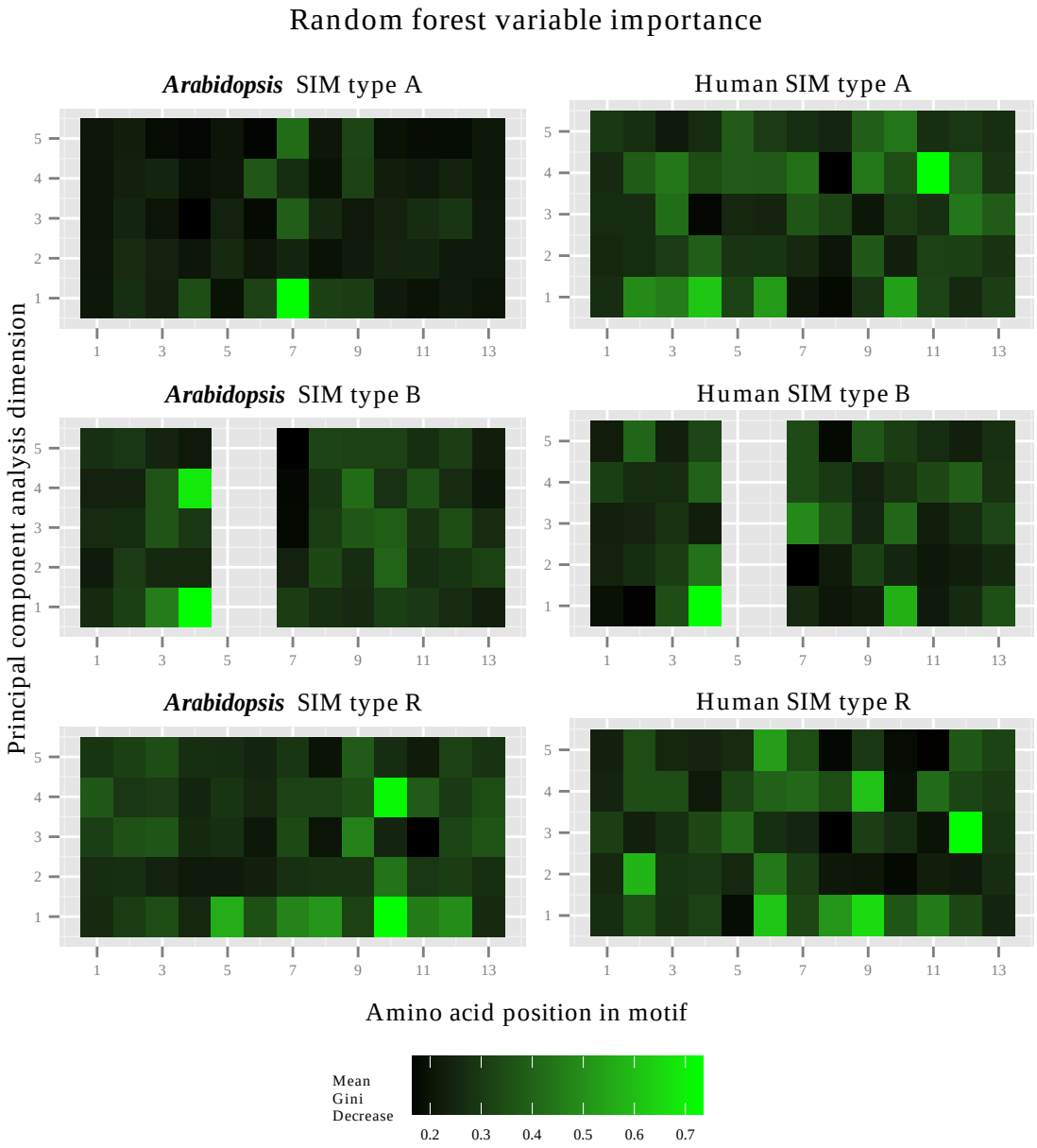


Figure 3.17: SIM predictor variable importance

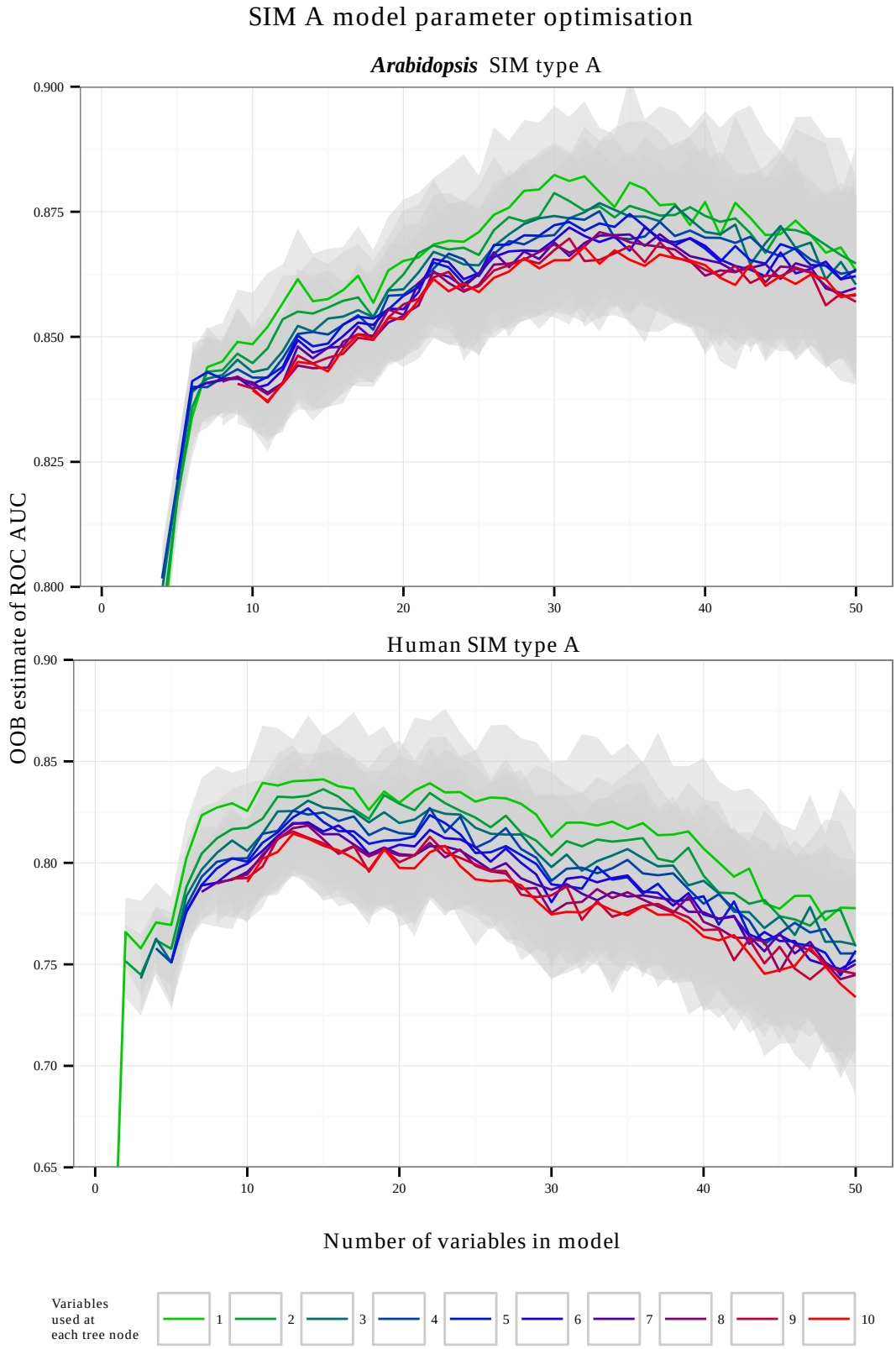


Figure 3.18: SIM A model parameter optimisation. Performance of random forest models was assessed by adding 1 variable at a time and testing a different number of variables sampled at each tree node. Shaded areas show the 95% confidence intervals, $n = 25$.

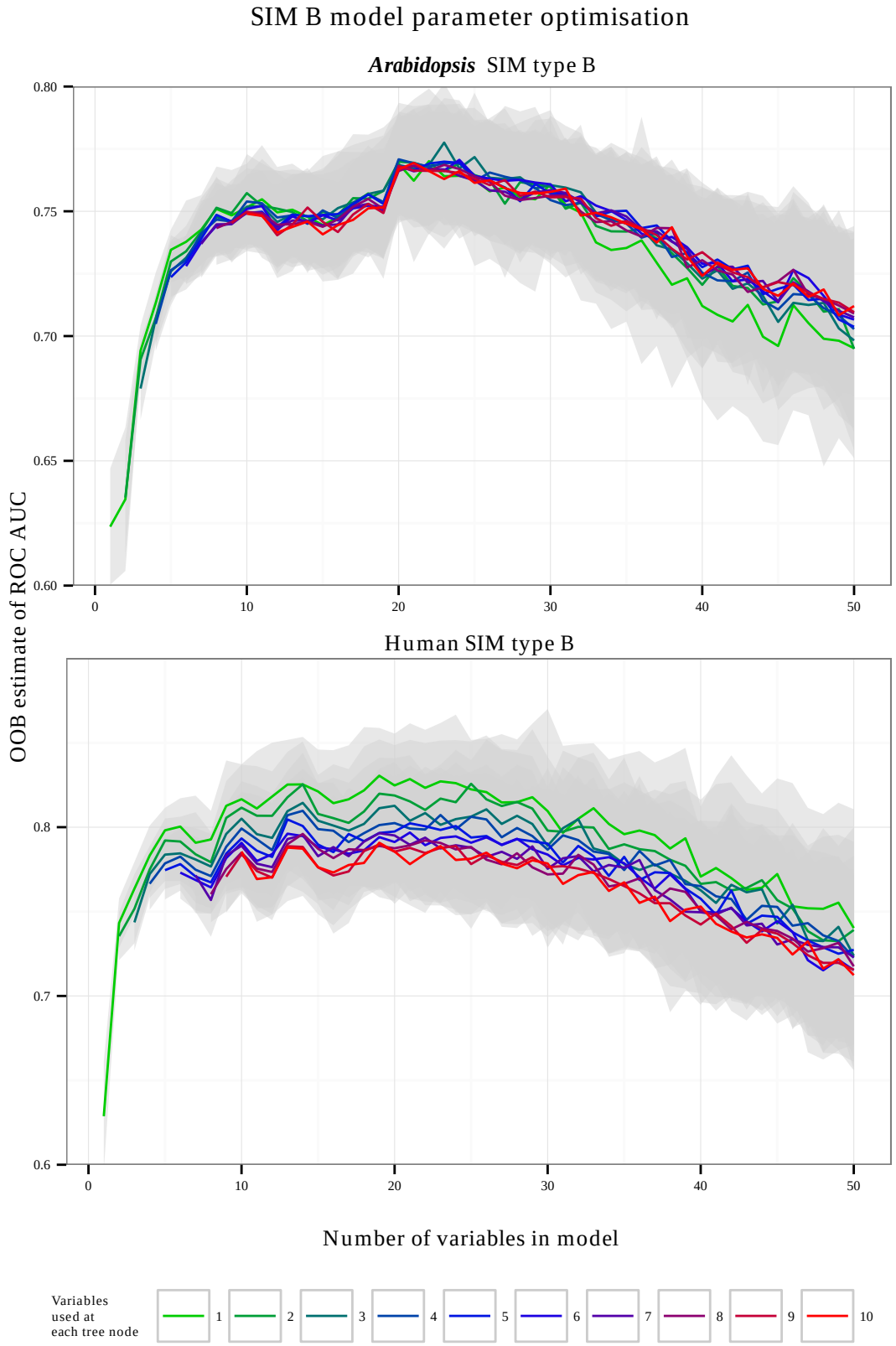


Figure 3.19: SIM B model parameter optimisation. Performance of random forest models was assessed by adding 1 variable at a time and testing a different number of variables sampled at each tree node. Shaded areas show the 95% confidence intervals, $n = 25$.

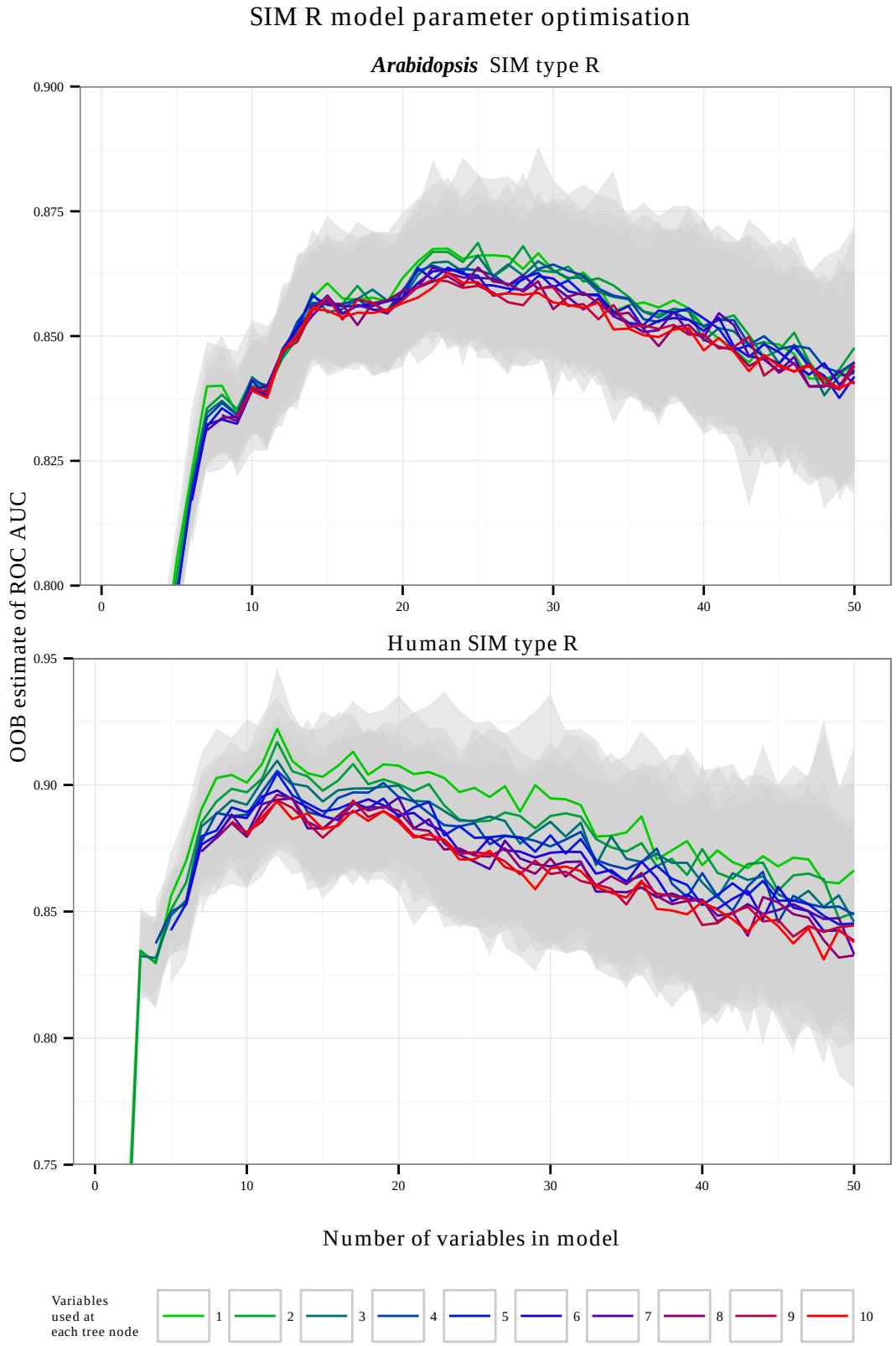


Figure 3.20: SIM R model parameter optimisation. Performance of random forest models was assessed by adding 1 variable at a time and testing a different number of variables sampled at each tree node. Shaded areas show the 95% confidence intervals, $n = 25$.

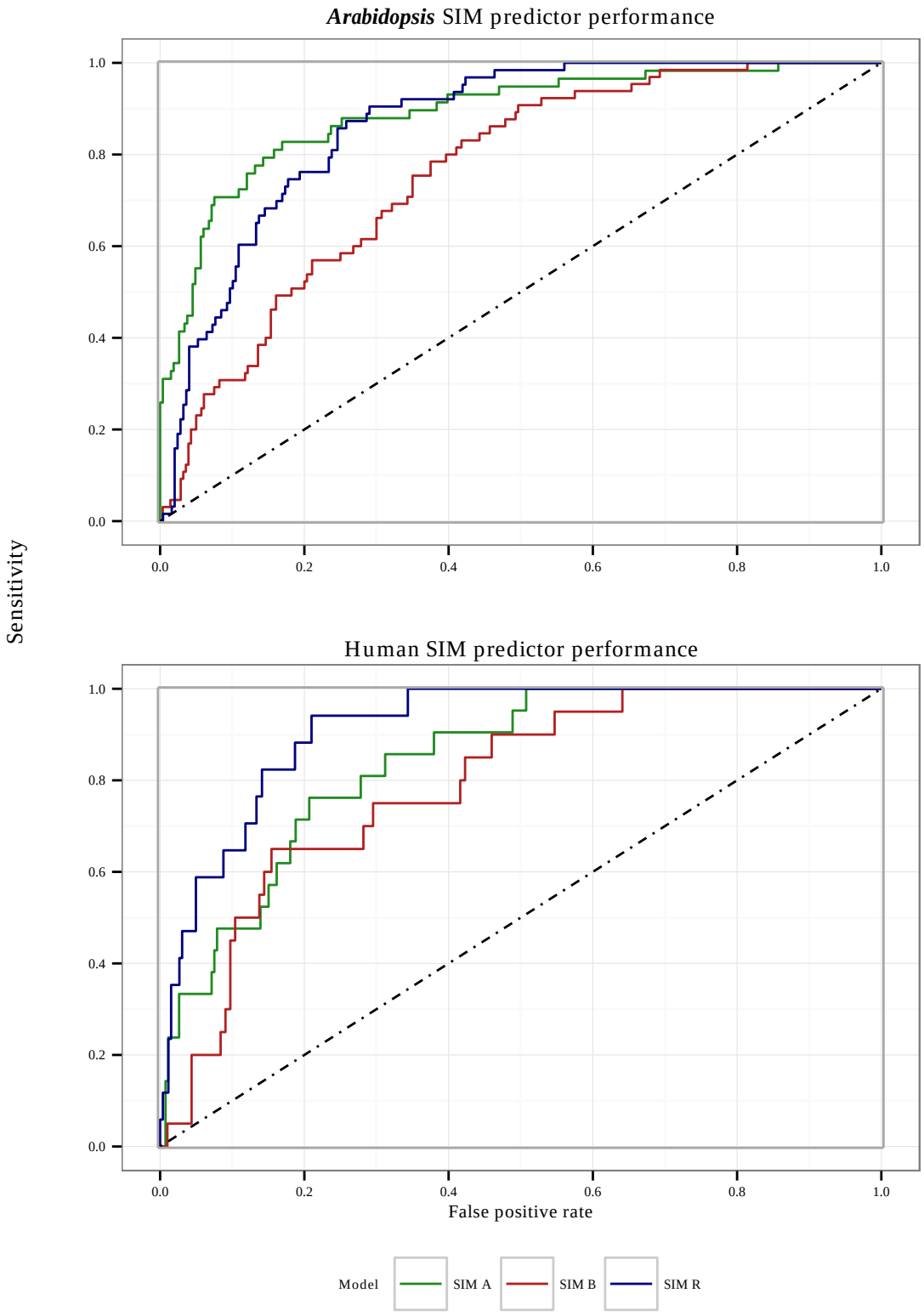


Figure 3.21: ROC curves of the performance of various SIM predictors.

SIM type	Species	ROC AUC (±95% CI)
A	*Arabidopsis*	0.887 ±0.0053
B	*Arabidopsis*	0.762 ±0.0118
R	*Arabidopsis*	0.869 ±0.0037
A	Human	0.850 ±0.0092
B	Human	0.817 ±0.0162
R	Human	0.920 ±0.0074

Table 3.4: SIM predictor AUC values. Mean ROC AUC values were obtained by repeatedly training random forests on the same data set. Random forests with 2000 trees were grown for each iteration. n = 25.

3.4.6 SUMO site random forest models

The SUMO site sequences were converted into 33 input variables for the random forest models. Unlike the SIM data, using 3 principal component dimensions rather than 5 resulted in better performance, suggesting that SUMO site prediction is less complex than SIM prediction. The differences between type I and II SUMO sites were notable and fit well with the understanding of these motifs, see Figure 3.22. For type I SUMO sites, the hydrophobic residue upstream of the central lysine is the most important followed by the [DE] feature at position 2, with other amino acids contributing a small amount of information. For the type II sites, the [DE] is most important but the strong hydrophobic feature upstream of the central lysine is absent, with the variable importance spread over the other positions. The central lysine contributes no information as sequences are preselected to have a lysine residue at this position so do not contribute any information to random forest predictors, thus no variables for this position are used in the predictor.

The optimisation algorithm was applied to the SUMO site data. As was found in the SIM optimisation, a performance maximum was found after which adding more variables reduced the performance of the model as measured by AUC. The performance drop however, was much less than for the SIM models. The optimal number of variables used at each tree node, m, was 1, the same as for most of the SIM models. This is in contrast to the random forest predictor by Teng *et al.* (2012) which was found to have optimal performance with $m = 6$, though their predictors used hundreds of input variables while the optimal number found in this work was 11 and 6 for type I and II respectively. For random forests using a higher number of variables, the optimal value for m starts to increase, this is especially apparent for the type I random forest with 33 variables in Figure 3.23 where $m = 1$ had the worst performance.

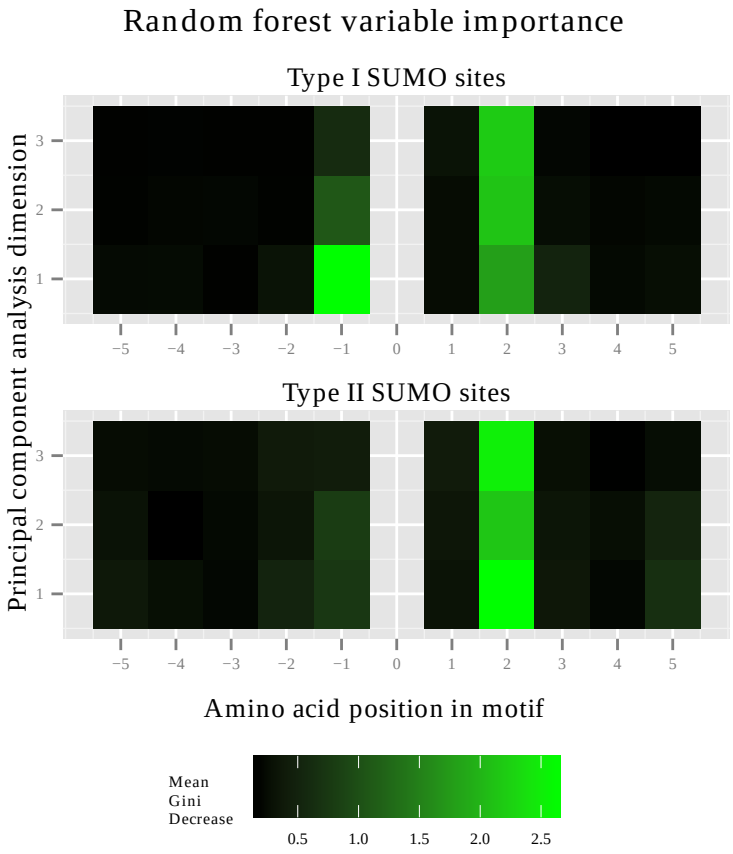


Figure 3.22: SUMO site predictor variable importance. The central SUMOylated lysine is at position 0 in the figure. Position 0 has no importance as sequences are prescreened to have a lysine residue at this position and so this position contributes no information to the prediction models.

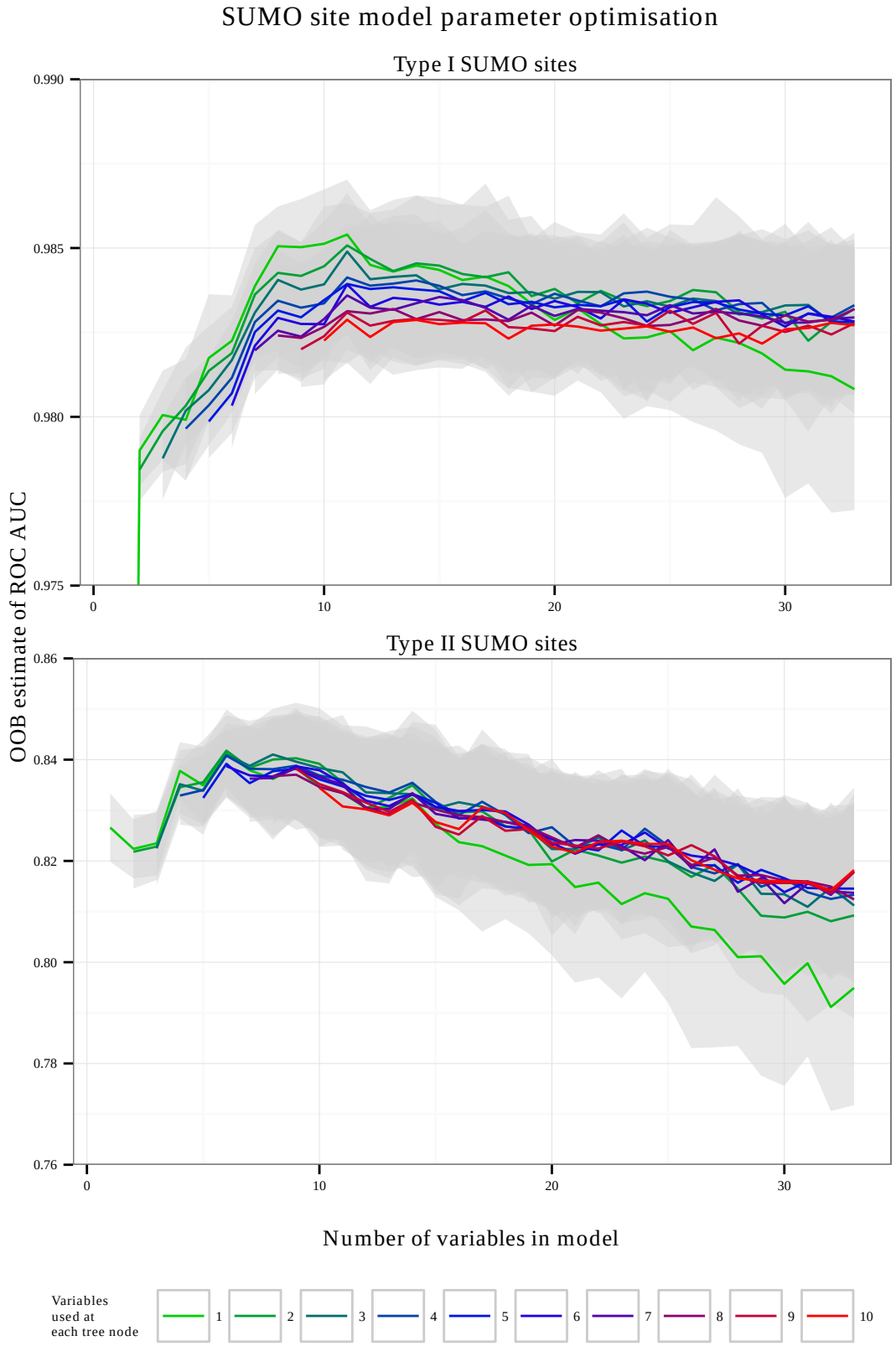


Figure 3.23: SUMO site model parameter optimisation. Performance of random forest models was assessed by adding 1 variable at a time and testing a different number of variables sampled at each tree node. Shaded areas show 95% confidence intervals, n = 25.

Once optimal parameters for the SUMO site predictors were identified, random forests were trained with 2000 trees, the point at which the OOB estimate of error could not be improved any further. For each SUMO site type, 25 random forests were trained and their performance was assessed by AUC values. Like the SIM random forest predictors, the variance of the AUC values between the random forests was very small. The performance was also compared with SUMOsp (Ren *et al.*, 2009) and seeSUMO (Teng *et al.*, 2012) by querying the training data used against these two predictors and using the resulting score values to calculate AUC values and ROC curves. The AUC for our model, known as HyperSUMO, and the other predictors are shown in Table 3.5 and the ROC curves are shown in Figure 3.24. The results show that as expected the type I predictor (AUC = 0.986) outperforms the type II predictor (AUC = 0.842) and the SUMO site predictors greatly outperform the SIM predictors (Table 3.4). The better performance of the SUMO site predictors is at least partly due to the much larger size of the training dataset but may also be influenced by the quality of the data and the complexity of the problem being addressed.

Model	SUMO types	ROC AUC (±95% CI)
HyperSUMO type I	I	0.986 ±0.00075
HyperSUMO type II	II	0.842 ±0.0039
SUMOsp type I	I	0.731
SUMOsp type II	II	0.725
seeSUMO	I & II	0.705

Table 3.5: Comparison of ROC AUC values for various SUMO site predictors. Mean ROC AUC for random forest models were obtained by repeatedly training random forests on the same data set (n = 25). Random forests with 2000 trees were grown for each iteration.

SUMOsp and seeSUMO were trained with the same data as was used to build the predictors and due to technical restrictions cross validation was not possible, therefore the resulting AUC values likely overestimate the performance of these predictors. Despite the possible overestimation of the performance of SUMOsp and seeSUMO, our model, HyperSUMO, greatly outperforms these models even for the less accurate type II predictor. The results for seeSUMO disagree with those published by the author who estimated a ROC AUC value of 0.920 for their best performing predictor while our results give a value of 0.705, which is an enormous discrepancy. One of the major differences between HyperSUMO's and seeSUMO's estimation of AUC is the validation dataset used, our method used all of the 8318 training sequences while Teng *et al.* (2012) used a separate set with a total of 1338 sequences of which 48 were SUMOylated. There may be something inherently different between these datasets that accounts for the discrepancy in estimated AUC values where the larger dataset gives a lower value and the smaller a higher value. Factors that could contribute to this difference include a different ratio of type I and II

sites or different accuracy of the data resulting from the methods used to identify the SUMO sites. The smaller dataset used for evaluation by Teng *et al.* (2012) was curated from publications after January 2010, whereas the data used in this work was from before this date.

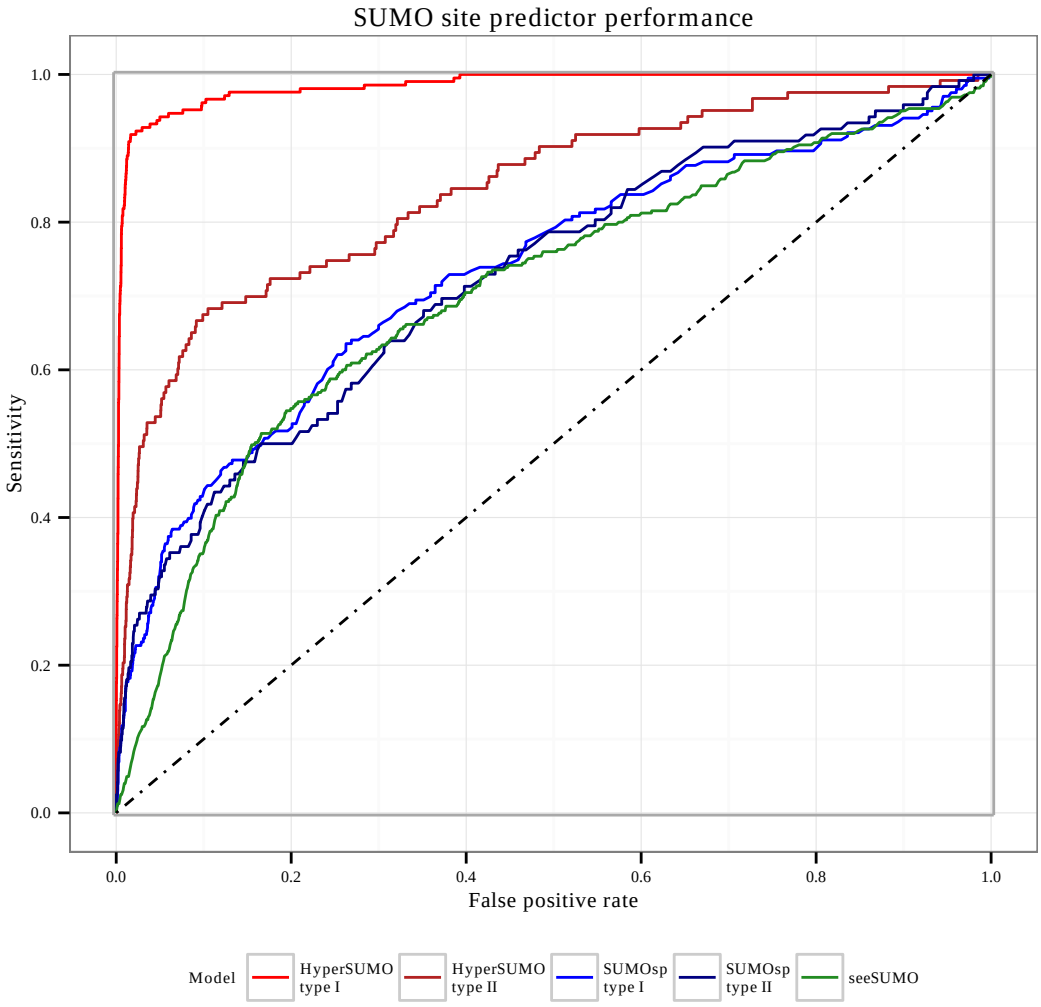


Figure 3.24: ROC curves of the performance of various SUMO site predictors. The published predictors SUMOsp and seeSUMO have comparable performance. HyperSUMO greatly outperforms the other two published predictors for both type I and type II sites, while the performance of type I prediction is markedly better. To calculate performance for HyperSUMO, FPR and TPR were calculated using OOB estimation using the full training dataset. For the other two models, the training data were queried against the predictors with the threshold set to 0 so that a score was generated for every lysine. SUMOsp and seeSUMO were trained with the same datasets but cross validation was not possible due to technical restrictions; therefore the ROC curves generated for seeSUMO and SUMOsp *over-estimate* their performance. Despite this HyperSUMO outperforms these predictors.

3.4.7 Genome screen for SIM containing proteins

A genome-wide screen for SIMs in *Arabidopsis* was performed which incorporated evolutionary information. Multiple orthologous genes from both eudicotyledonous and monocotyledonous plant species were aligned with *Arabidopsis* genes and conserved SIM-like motifs were identified that were outside

of predicted protein domains. The top 500 identified genes are listed in Table B.2 in the Appendices. GO term analysis revealed that this set of genes was enriched for cell wall and sugar metabolism, DNA/RNA homeostasis, transport and calmodulin binding. These results are in accordance with the known role of SUMO in chromatin remodelling and a number of notable genes that were identified include *ATDDM1*, *CHR8* and *CHR31*. A summary of the molecular function GO terms is shown in Table 3.6 and biological process GO terms are shown in Table 3.7. The full list of GO term results are shown in Tables B.3 and B.4 in the Appendices.

Along with the DNA/RNA maintenance genes, a number of DNA repair genes were identified including *RAD5*, *GMI1* and *ATXPD,* and a wide range of stress genes including heavy metal induced genes and heat shock protein genes were also identified. This suggests a strong role for SUMO in coping with environmental stress and DNA damage, roles which have been cited in the literature (Mazur & van den Burg, 2012; Castro *et al.*, 2012).

Molecular function GO term	**Observed frequency (%)**	**Expected frequency (%)**	**Ratio**	***p*-value**
Cellulose synthase activity	4.09	0.57	7.24	2.59E-08
Calmodulin binding	2.50	0.15	17.09	7.04E-06
ATP binding	11.36	3.01	3.78	4.26E-05
Adenyl ribonucleotide binding	11.36	3.03	3.75	4.49E-05
ATPase activity, coupled to transmembrane movement of ions, phosphorylative mechanism	11.14	2.87	3.87	4.62E-05
Purine nucleoside binding	3.64	0.70	5.23	4.62E-05
Beta-galactosidase activity	11.14	2.99	3.72	4.90E-05
Helicase activity	9.55	3.55	2.69	1.59E-04
Ribonucleotide binding	20.00	11.62	1.72	2.71E-04
DNA-dependent ATPase activity	12.05	4.97	2.42	5.76E-04
Nucleotide-sugar transmembrane transporter activity	8.18	3.05	2.68	8.17E-04
Clathrin binding	6.59	1.70	3.89	2.05E-03
Hydrolase activity, acting on acid anhydrides, catalysing transmembrane movement of substances	5.00	1.26	3.96	2.05E-03

Table 3.6: Summary molecular function gene ontology analysis.

Biological process GO term	**Observed frequency (%)**	**Expected frequency (%)**	**Ratio**	***p*-value**
Plant-type cell wall biogenesis	2.74	0.20	13.68	1.10E-07
Cellular developmental process	5.94	1.51	3.92	1.14E-06
Cell morphogenesis involved in differentiation	3.42	0.56	6.06	6.11E-06
Transport	14.16	6.94	2.04	8.01E-06
DNA-dependent DNA replication initiation	1.37	0.04	31.08	1.47E-05
Carbohydrate biosynthetic process	4.11	0.96	4.29	2.80E-05
Organelle organisation	6.16	2.16	2.85	9.98E-05
Polysaccharide biosynthetic process	2.51	0.39	6.40	1.21E-04
Disaccharide metabolic process	1.83	0.19	9.70	1.71E-04
Embryo development	5.02	1.65	3.04	2.77E-04
Cellular carbohydrate metabolic process	5.25	1.80	2.91	3.27E-04
Oligosaccharide metabolic process	1.83	0.23	8.00	4.87E-04
Cell growth	3.65	1.01	3.63	5.46E-04
Response to light stimulus	5.25	1.91	2.75	6.64E-04
Embryo development ending in seed dormancy	4.34	1.42	3.05	8.05E-04
Cellulose biosynthetic process	1.37	0.12	11.03	9.09E-04
Intracellular transport	4.57	1.57	2.90	9.09E-04

Table 3.7: Summary biological process gene ontology analysis.

3.5 Discussion

3.5.1 Peptide array

The use of arrays of synthetically generated peptides to screen for SUMO interacting peptides was successful and the models used to generate those peptides produced a good balance of interacting and non-interacting peptides, with about 20% or 10% of the total number of peptides interacting with AtSUM1

or HsSUM1 respectively. Had the peptide generation models been too specific, too few non-interacting peptides would have been identified; conversely, had the models been too general, too few interacting peptides would have been identified.

There were a number of issues with the peptide arrays and the methods used to collect interaction data. There was a large variance in the amount of peptide produced in each spot and the lack of a suitable method to accurately measure the peptide amount resulted in uncertainty in the amount of peptide present which prevented any quantitative analysis of intensity signals for the peptide spots. Peptide amounts could not be accurately quantified due to the limitation of using dye-based methods for short peptides, where the intensity of the staining varies greatly depending on the amino acid composition of the peptides. A similar difficulty is experienced with spectroscopic methods, where the aromatic amino acids have a higher absorbance in the UV spectrum used for quantification (Hansen *et al.*, 2013). For larger protein molecules, dyes or spectroscopy are accurate methods of quantification as the proportions of amino acids between proteins is usually similar. However, for short peptides the proportions of amino acids vary significantly between peptides. It may be possible to develop a more accurate method of quantification if the sequence of the peptides being measured is known, which is the case for peptide arrays. For UV spectroscopy the absorbance of the different amino acids is well characterised. The sequence of amino acids in a peptide could be used to estimate an absorbance per quantity of peptide. Such a method could then be used to accurately quantify the amount of peptide in array spots and differences in interaction strength could be estimated with higher accuracy.

A confounding issue with synthetic peptide arrays is the purity of the resulting peptides as the peptide spots are synthesised on the array, leaving any mis-formed or truncated polymers and often, as was the case with this work, the purity of the peptides is not assessed. However, work by Frank (2002) has shown that generally peptide spots on arrays are of high purity. Given the large number of peptides in an array, it is likely that at least a small number will be of low purity due to variation in the synthesis method and synthesis of difficult peptide sequences. Data generated for sequences with low purity will be inaccurate as impurities may inhibit a true interaction or produce a false interaction; also if the amino acid sequence is used to normalise peptide amount measurements, low purity will lead to inaccurate results. Given the likelihood that at least some peptides in the arrays used in this work will have low purity, it is likely that a number of the interaction results are false though as long as the number of these false results is low, both analysis of the binding peptides and the prediction models should only be affected to a limited degree since RF models have been shown to be very tolerant to misclassified data (Breiman, 2001).

Western blotting was used to detect peptide interactions and significant difficulty was encountered with HsSUM1 as the antibody used had high cross-reactivity resulting in a large number of the peptide spots having to be excluded from analysis. The source of the cross-reactivity is not known and may

have come from either the control GST protein binding to the peptide spots or the antibodies themselves binding directly to the peptides. Although different antibodies were used to detect the HsSUM1 interaction, no combination was found that did not exhibit this cross-reactivity issue. Cross-reactivity in far-western blotting is a frequently occurring issue and arises due to the large number of different steps in the procedure that can generate non-specific binding. These include offsite binding of the probe protein, binding of a protein tag on the probe protein if present and binding of either the primary of secondary antibody to the peptides (Katz *et al.*, 2011). Using different methods, with fewer steps, can alleviate this issue with fluorescently labelled probe proteins being a suitable alternative. Traditionally probe proteins were labelled by non-specifically attaching a fluorescent molecule. This method had two major drawbacks though: the sites of probe labelling varied between protein molecules and it was difficult to achieve consistent labelling between labelling reactions. The stochastic nature of the labelling can also lead to important sites of the probe protein being modified inhibiting the interaction between the probe and target. Newer methods such as using fluorescently labelled amino acids during protein translation or using site specific labelling methods have alleviated these issues. The HaloTag® system from Promega can be used to specifically label a target protein. A HaloTag® plasmid is used to express the probe protein fused to a HaloTag® protein. The HaloTag® protein can then be modified with a number of fluorescent ligands that the tag specifically recognises and covalently attaches to, leaving the probe protein unmodified. The HaloTag® however is very large and there is the possibility that its large size may interfere with the interaction between the probe and its target, though fusing the tag to the opposite terminal may alleviate this problem (Hurst *et al.*, 2009).

Fluorescently labelled probe methods additionally have the advantage that they are quantitative, allowing the strength of a protein interaction to be measured. Far western blotting on the other hand is generally semi-quantitative due the non-linear response of the chemiluminescent detection and the number of binding steps, though methods described by Weiser *et al.* (2005) have been able to produce quantitative results from peptide arrays using far-western blotting by calibrating the method with a series of interactors with known interaction strengths. Overall though, future work would probably benefit from using a fluorescently labelled probe to accurately measure the relative interaction strength combined with accurate measurement of peptide amounts in the arrays to normalise the intensity results. Having accurate interaction strength results would allow more accurate models of SIM binding to be constructed so a distinction between strong and weakly binding SIM peptides could be made. Actual interaction strength results would allow the use of regression models rather than the classification models used in this work, which would allow the prediction of actual interaction affinities.

3.5.2 SIM analysis

The SIM library was used to screen the interaction of around 1000 peptides with AtSUM and around 800 for HsSUM1. This is the first investigation that compares isoforms from two distinct organisms and is also the first to screen such a volume and diversity of peptides, making it the first comprehensive comparison of SUMO isoform binding preferences. Namanja *et al.* (2012) used a similarly large number of peptides to characterise the binding of two human SIMs against HsSUM1 and HsSUM3, and the technical data from their work was crucial for the design and implementation of the SIM library described here. Overall the results agree well with other published work which has demonstrated that, depending on the amino acid composition, SIM peptides either bind specifically to one SUMO isoform (Cai *et al.*, 2013; Sekiyama *et al.*, 2008) or can be more general and bind to many isoforms within a species (Escobar-Cabrera *et al.*, 2011). The emerging paradigm is that there are a diverse array of SIM motifs within proteins with different binding specificities for the paralogous SUMO isoforms within a species and these differences allow the SUMO isoforms to perform distinct roles within the cell. Thus the outcome of SUMOylation of a target can depend on the isoform it was modified with. Further specificity is achieved with recognition proteins containing both SIMs and target protein specific interacting domains (Armstrong *et al.*, 2012) allowing SUMOylation to have specific functions for different proteins.

Chain topology is another critical factor in SIM interactions, which was not explored in this work. Often SIMs are found in tandem in proteins, with each motif interacting with a SUMO residue within a poly-SUMO chain, and it has been shown that in some cases poly-SUMO chains rather than a single SUMO are required for interaction (Tatham *et al.*, 2008). It is possible that chains of mixed SUMO isoforms may also have specific functions and investigating the role of SUMO chain topology on protein regulation would be an interesting topic for further research.

Using the library of SIM peptides, amino acid preferences along the SIM peptide sequences were characterised and the binding of the AtSUM1 and HsSUM1 isoforms were compared, which as expected showed vastly different binding preferences with only a small number of peptides able to bind to both SUMO isoforms. The differences in the SIMs that resulted in isoform selection were found to be due to amino acids flanking the hydrophobic core, with the charge of these positions having a strong effect. Analysis of HsSUM1 binding preferences was hampered by the small number of positive instances of interaction which was due to a large number of peptides being excluded from analysis due to non-specific antibody binding and due to a smaller proportion of positive interactions in the test set (10% for HsSUM1 compared to 20% for AtSUM1).

Post translational modification of SIM peptides and SUMO adds an addition layer of complexity and phosphorylation has been shown here to be a powerful regulator of the SIM-SUMO interaction. There are a number of examples of SIM phospho-switches described in the literature in human systems that demonstrate crosstalk between SUMOylation and protein kinase cascades (Percherancier *et al.*, 2009;

Stehmeier & Muller, 2009), however, the extent to which phosphorylation plays a role in plants remains unknown as no examples of SIM phospho-switches have been described and the data presented here shows that the interaction promoting effects of phosphorylation are restricted in AtSUM1 interactions. Charged amino acids are over-represented in regions flanking the hydrophobic cores of SIMs, with negatively charged groups very common. These charged groups play a role in stabilising the SIM peptide within the SUMO groove and it is plausible to speculate that phosphorylation of polar residues within SIM enhances SUMO binding by playing a role similar to that of the negatively charged amino acids. Additionally phosphorylation imparts a -2 negative charge, twice as much as the negatively charged amino acids, suggesting that it can have a strong effect on the electrostatic attraction of a SIM to SUMO. The positioning of positively charged amino acids with the two SUMO isoforms studied is different, with HsSUM1 containing more positively charged groups within the important α-helix 1 and β-strand 2. It is possible that the increased number of positively charged groups in HsSUM1 were responsible for the higher proportion of phosphorylated interacting SIMs observed, compared to AtSUM1 where a very large proportion of interactions were abolished when the SIM peptides were phosphorylated.

Immobilised recombinant tandem SIMs have been used to selectively purify SUMO and SUMOylated proteins from total cell lysates and have been shown not to bind to similar ubiquitin-like proteins, including ubiquitin itself (Da Silva-Ferrada *et al.*, 2013). The ability to specifically purify SUMOylated proteins can be used to identify which proteins are SUMOylated under specific conditions and expand the understanding of SUMO regulated biochemical pathways. The work shown here could hopefully be used to extend these SUMO purification methods by designing short peptides with higher affinity to SUMO or to design peptides with specific affinity to different SUMO isoforms within a species. Being able to specifically purify a particular SUMO isoform would be useful in studying the divergent roles of the different isoforms.

3.5.3 SUMO sequence feature predictors

The SIM data and SUMO site data from previous publications were used to build SUMO sequence feature predictors using random forest models. The performance of the SUMO site predictors as measured by AUC values was much better than the SIM predictors. The difference between the two is probably due to the extensive dataset for SUMO sites and smaller dataset sizes for the SIM data. The SIM features were also more complex (i.e. more amino acids were important for determining an interaction) and this was exacerbated by the limited dataset size. Nevertheless the SIM predictors have a reasonable performance and are the first example of SUMO isoform specific predictors.

One of the major aims of this work was to optimise the performance of the predictors using variable selection and parameter optimisation. Random forests had been used earlier by Teng *et al.* (2012) to

predict SUMO sites so their model, which did not utilise variable selection to an appreciable degree, was a useful benchmark to compare with. PCA was used to reduce the multitude of amino acid indices into a small number of variables which were used to convert amino acid factors into numeric vectors. Using numeric variables was shown by Teng *et al.* (2012) and this work to improve the performance of random forest predictors, however, Teng *et al.* (2012) did not use a PCA decomposition of the variables but rather selected a number of useful amino acid indices instead. Through the course of developing the random forest predictors, it was found that the optimal variables to use came from converting different amino acid positions into different combinations of PCA dimensions, rather than using the same dimensions for all amino acids positions in the predictor. For example, position two in a peptide may be converted into the first dimension while position three may be converted into the second and third. Though this usage of variables was rather complicated and added complexity to data processing in the predictors, the performance of the predictors significantly benefitted. Compared with previous work, the SUMO site predictor developed here had significantly better performance. The performance increase was so large that initially there was concern that an error had been made in assessing the performance of the other published predictors, however, the same set of peptides was used to assess all predictors and OOB sampling was used for our predictor which controls for over-estimation of performance. Use of OOB was not possible for assessing the other published predictors, seeSUMO and SUMOsp, but if anything this should slightly overestimate the performance of these predictors as they used similar training data to the data used to assess them.

The method of carefully optimising random forest variables and using PCA decomposition of factor variables has shown that significant improvements can be made to predictor performance. Optimisation and removal of redundant data reduces the complexity of the input data and it is likely that optimisation of these factors was responsible for the observed increase in predictor performance. With the large body of data available for SUMO sites and the high performance of the predictors for these features, it is unlikely that large improvements in predictor performance can be achieved in future work, at least for analysing primary sequence data alone. However, the SIM prediction models could probably be improved significantly if more data were generated to train the predictors. Also, including SIM binding data for other SUMO isoforms in important research organisms is a priority as there are four different functional SUMO isoforms in both human and *Arabidopsis* and one in yeast. It would also be beneficial to generate quantitative data on the affinity of the different SUMO isoforms as predictors could be used to identify the interaction strength of a given peptide. One limitation of the work presented here is that the peptide sequences used in the arrays were highly constrained to be similar to known SIM peptides. This approach was used to ensure that sufficient results would be obtained from the limited number of peptides screened, however, due to the constraints imposed, any binding peptides that were outside of these constraints would have been missed and the features responsible for their interaction not

captured. If a sufficiently large number of peptides could be screened, it would be beneficial to lower the constraints imposed on peptide design to sample a more diverse range of peptides. Such peptides could be constrained to only contain the essential hydrophobic cores while allowing every other position in the peptides to vary freely. It is likely that this would result in fewer positive hits but may also detect any peptide features missed in this work.

3.5.4 *Arabidopsis* genome-wide SIM screen

The SIM prediction models were used to predict putative SIMs within the *Arabidopsis* genome. To increase the specificity of the screen, structural and evolutionary information was used to inform the results. Since SIMs are unstructured regions of protein they tend to lie outside of functional domains and any predicted SIMs lying within regions predicted to be functional domains using the Conserved Domains Database (Marchler-Bauer *et al.*, 2013) were excluded. Predicted SIMs also had to be present within alignments of orthologous proteins from diverse species extending to the monocots to fulfil the assumption that functionally important SIMs would be conserved in a diverse range of species. *Arabidopsis* proteins for which orthologs were not identified were removed so any narrow, clade specific genes resulting from gene duplication events were removed. The constraints imposed in this screen were targeted towards detecting conserved, functionally important SIMs and would not identify any functional SIMs that evolved more recently within a limited clade of plant species. To control the FPR, the cutoff thresholds for various variables in the search were high and it is likely, especially given the suboptimal sensitivity of the SIM predictors, that many genuine SIMs will have been missed by this screen although hopefully the genes identified in this screen will provide an initial step in understanding the genome-wide role of SUMO binding proteins (SBPs) in plants. Recent genome-wide screens for SUMOylated proteins (Miller *et al.*, 2010, 2012) and ubiquitinated proteins (Maor *et al.*, 2007) have been successful in identifying proteins that are modified by these two ubiquitin-like proteins and the next step in unravelling the role of SUMOylation will be to identify the fate of those modified proteins.

The highest scoring 500 genes identified in the screen were selected as being likely to contain at least one functionally relevant SIM. No estimate of the FPR for these results could be determined and subsequent validation with laboratory experiments is required. However, gene ontology analysis identified functional enrichment in a number of categories for this gene set. The genes were functionally enriched in three broad areas: DNA/RNA maintenance, sugar and cell wall metabolism and transport with the DNA/RNA maintenance agreeing well with published literature on the role of SUMOylation (Mazur & van den Burg, 2012; van den Burg *et al.*, 2010). Interestingly, the identified genes do not show nuclear localisation enrichment which is reported widely in the literature for SUMOylated proteins, rather these predicted SIM containing proteins showed enrichment for membrane bound localisation (both plasma and organelle membranes) and cytoplasmic localisation. This localisation enrichment is due to the high

number of biosynthetic and cell wall related genes in the set.

The DNA/RNA related set of genes comprise a large number of nucleotide binding proteins implying an important role in both DNA and RNA regulation. The genes *DDM1*, *RAD5*, *GMI1* and *UVH6* (*AtXPD*) are distinct DNA repair related genes that were identified. All of these genes play a role in repairing DNA damage resulting from oxidative damage, ultraviolet radiation or gamma radiation and play an important role in oxidative stress tolerance. RAD5 and GMI1 are both involved in repairing DNA double strand breaks and appear to do this through homologous recombination (HR) as knock downs of these genes alter HR patterns although the exact mechanism is unknown (Chen *et al.*, 2008; Böhmdorfer *et al.*, 2011). UVH6 is an ATP dependant helicase that unwinds damaged DNA in preparation for repair by DNA repair components and this protein also appears to play a role in development as knock downs show defects in floral development as well as sensitivity to DNA damaging agents and heat (Ly *et al.*, 2013). DDM1 is a cytosine methyltransferase enzyme that plays a role in chromatin remodelling and gene regulation (Ogrocká *et al.*, 2014) and is important for DNA repair as knock downs are sensitive to oxidative DNA damage though the mechanism for sensitivity is as yet unknown (Qüesta *et al.*, 2013). CHR8 and CHR13 were another two chromatin remodelling enzymes with helicase activity that were identified and have been also been implicated in DNA repair (Shaked *et al.*, 2006). Such a large number of diverse DNA repair related proteins identified is suggestive of the DNA repair process being regulated by SUMOylation and it is likely that these proteins are the interacting partners of SUMOylated proteins that have previously been implicated in DNA repair (Saracco *et al.*, 2007).

Two 90 kDa heat shock proteins (HSPs) were identified fitting with the theme of the SBPs playing a role in oxidative stress. 90 kDa heat shock proteins are chaperones which are required for both normal protein folding and for refolding of denatured proteins. The two identified HSPs have low sequence identity and appear to have different functions: Hsp90.4 is constitutively expressed and likely plays a role in protein homeostasis under normal conditions while Hsp90.1 is stress-induced and probably plays a more dominant role in protein protection and repair during stress conditions (Cha *et al.*, 2013).

The strong enrichment for cell wall and sugar biosynthesis proteins suggests a novel role for SBPs and there are very few examples of SUMO playing a direct role in these functions in the literature. Miura *et al.* (2011) found that the SUMO E3 ligase SIZ1 regulated root growth and architecture and mutant *siz1* plants differentially expressed genes encoding cell wall loosening and cell wall biosynthesis genes which were under the control of auxin. Interestingly, the SBP screen identified one auxin receptor protein, AFB3 (Parry *et al.*, 2009), and two auxin efflux carrier proteins, PIN1 and AGR, with the latter expressed solely in the root (Petrásek *et al.*, 2006). It is tantalising to speculate that these auxin related proteins are directly regulated by some as yet unknown SUMOylated protein or proteins and that this regulation is responsible for the reduced root growth phenotype observed during hyper-SUMOylation under stress (Conti *et al.*, 2008).

Chapter 4

Characterisation of SUMOylated RGA

4.1 Introduction

One of the major phytohormones responsible for growth regulation in plants is the gibberellin (GA) group of hormones which regulate growth in response to a multitude of environmental cues and play an important role in developmental processes (Schwechheimer & Willige, 2009). On a molecular level the GAs regulate the stability of a group of transcriptional repressor proteins known as DELLAs which are degraded by the GA receptor GIBBERELLIN INSENSITIVE1 (GID1) in the presence of GA (Ueguchi-Tanaka *et al.*, 2005). The primary role of the DELLA proteins in the GA pathway is to restrain growth and developmental processes regulated by GA by inhibiting a diverse range of transcription factors. In the presence of GA, the DELLA proteins are degraded and the repression on the transcription factors they inhibit is released.

The DELLA proteins are nuclear localised and consist of two functional domains. The N-terminal region of the protein consists of a 'DELLA' domain, so called for a conserved motif with this sequence. The DELLA domain of the protein is recognised by GID1 and is responsible for regulating the stability of the protein. The C-terminal region contains a GRAS domain which is found in a large number of protein classes (Bolle, 2004). The GRAS domain forms interactions with other proteins and is mostly responsible for the repressive effects of the DELLA proteins (Sun & Gubler, 2004). While the dominant role of the DELLA proteins is to repress transcription factors, the DELLA domain of the protein is also a transcriptional activator that positively regulates a number of genes (Hirano *et al.*, 2012). The DELLA proteins regulate a diverse range of transcription factors that regulate processes including growth, floral and seed development and crosstalk with other hormone pathways (Locascio *et al.*, 2013).

Recent work on the SUMO protease double mutant plants *ots1 ots2* first characterised by Conti *et al.* (2008) lead to the discovery that the *Arabidopsis* DELLA proteins RGA and GAI were SUMOylated and the SUMOylation of these proteins was elevated under stress conditions. Interestingly, plants exposed to salt stress show a reduced growth phenotype that is not explained by GA levels alone, but rather it was found that the DELLA proteins accumulate under the salt stress conditions, despite the presence of GA which would normally lead to their degradation (Conti *et al.*, 2014). Initially it was suspected that SUMOylation of the DELLA blocked GID1 mediated degradation, however, only a small pool of the DELLA proteins were ever SUMOylated, which could not explain the accumulation of the unmodified DELLA proteins. Conti *et al.* (2014) proposed a model whereby SUMOylated DELLA proteins bind to GID1 but are not degraded, and by binding to GID1 the SUMOylated DELLA proteins sequester the receptor, preventing it from targeting the pool of unmodified DELLA proteins for degradation thus acting as an inhibitor of GID1. The ability of SUMOylated DELLAs to bind to GID1 was shown using co-immunoprecipitation of SUMOylated RGA with GID1a and the interaction with the SUMOylated DELLA protein was shown to be independent of GA. The results presented by Conti *et al.* (2014) showed the GA pathway in plants can be regulated independently of GA and shows that there is crosstalk

between the GA pathway and the SUMO cascade process, revealing previously unknown complexity in the regulation of plant growth and development.

A number of questions remain open with regards to the mechanism of protection of DELLA proteins by the small pool of SUMOylated DELLA proteins and although a model of SUMOylated DELLA proteins sequestering GID1 has been proposed, there is as of yet no direct evidence of the mechanism and a number of alternative processes could be responsible. One of the most important requirements for the sequestering of GID1 by SUMOylated DELLAs is that the binding affinity of the two molecules is strong enough to allow a small pool of SUMOylated DELLA to have such a large inhibitory effect. Additionally it would be important to test if there was any difference in the affinity between the hormone bound GID1 complex and unbound GID1. To answer these questions, quantitative data on the interaction kinetics is required to test this model and the results would inform the direction of future research into the GA pathway.

The original goal of the work presented in this chapter was to investigate the binding kinetics of GID1a with SUMOylated RGA using surface plasmon resonance (SPR) to generate the data required to test the model proposed by Conti *et al.* (2014). However, the task of producing sufficient amounts of RGA proved to be technically challenging and rather this chapter presents methods to produce purified RGA protein and an *in vitro* cell free method to SUMOylate the protein which may be used in work to investigate the binding kinetics with GID1.

4.2 Chapter aims

- Find the site of RGA SUMOylation
- Produce and characterise SUMOylated RGA protein

4.3 Results

4.3.1 Analysis of RGA protein sequences

Previous work had shown that the DELLA proteins RGA and GAI were SUMOylated (Conti *et al.*, 2014; Nelis, 2011) but the site of modification was not identified. In order to explore the role of SUMOylated RGA and other DELLA proteins, the SUMOylated lysine needed to be identified in order to generate a non SUMOylatable mutated version of the protein and to analyse the structural consequences of SUMOylation.

All five *Arabidopsis* DELLA proteins and various DELLA proteins from other species were aligned and the peptide sequences were analysed with SUMOsp 2.0 (Ren *et al.*, 2009) to identify putative SUMO sites (Figure 4.1). A conserved lysine in the DELLA domain of the protein was identified which was predicted to be a SUMO site in all protein orthologs, corresponding to K65 in RGA or K49 in GAI. This was the only lysine within all the DELLA proteins that was consistently predicted to be a SUMO site and based on the structure of GAI was surface exposed, making it the best candidate lysine for further analysis.

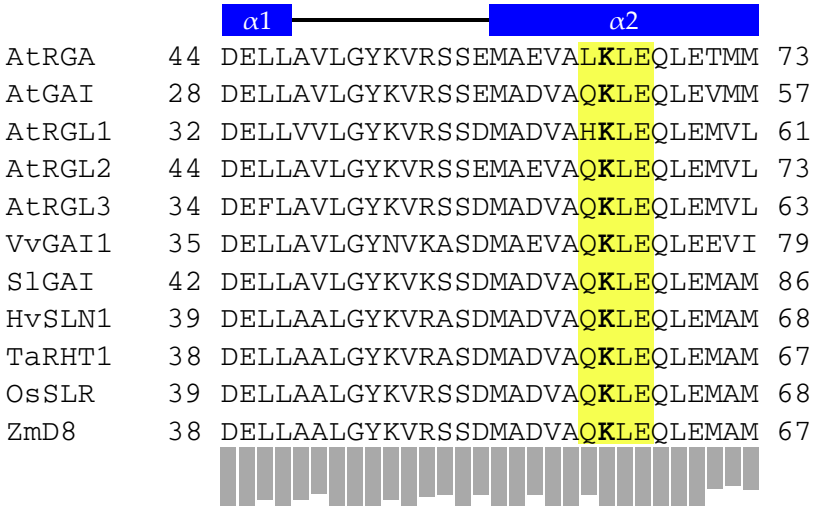


Figure 4.1: Alignment of DELLA proteins showing the predicted SUMO site within the conserved DELLA domain. DELLA proteins from a wide evolutionary range of species show strong conservation (grey bars below alignment). The predicted SUMOylatable lysine lies within α helix 2 of the protein, within the domain that interacts with GID1. The location of the alpha helices within the protein are shown in blue and are based on the crystal structure of AtGAI from Murase *et al.* (2008). Species used: Vv = grape, Sl = tomato, Hv = barley, Ta = wheat, Os = rice, Zm = maize.

The predicted SUMO site lysine lies within a highly conserved region of the DELLA protein, which shows very few sequence differences even in monocotyledonous plants, suggesting functional importance for this region. The DELLA domain of the DELLA proteins forms the surface that interacts with the GID1 GA complex and the identified lysine lies within the α-helix 2 of the protein which both makes contact with GID1 and forms a salt bridge with α-helix 4 within the protein (Murase *et al.*,

2008). SUMOylation of this lysine would almost certainly prevent DELLA-GID1 binding through steric hindrance of binding conformation, suggesting that SUMOylation of the DELLA proteins inhibits direct interaction with hormone bound GID1.

4.3.2 Lysine 65 in RGA is the site of SUMOylation

Based on bioinformatic analysis of multiple DELLA protein sequences, lysine 65 in AtRGA was a strong candidate as the site of SUMOylation. To determine if this was indeed the site of SUMOylation and the only SUMO site within the RGA, *RGA* and *rga K65R* clones in the plasmid pENTR were acquired from a collaborator to test SUMOylation at this site. N-terminal His tag fusions were generated by recombination of the pENTR clones with pDEST17 and these were transformed to the *E. coli* reconstituted SUMOylation system (Okada *et al.*, 2009) using *SAE1a* + *SAE2* as the E1 heterodimer, *SCE1* as the E2. Two forms of AtSUM1 were used, an active form (SUM1-GG) and a form with the terminal diglycine mutated to diarginine (SUM1-AA) which cannot be ligated to other proteins as a negative control for the assay.

The RGA proteins were expressed in the reconstituted system for 2 hours at 30ºC and then total cell lysates were analysed by western blotting using αRGA antibodies to visualise the proteins. The assay (Figure 4.2) showed that in the reconstituted *E. coli* system, rga K65R was not SUMOylated while the wild type protein was showing that lysine 65 is the SUMOylated residue in RGA.

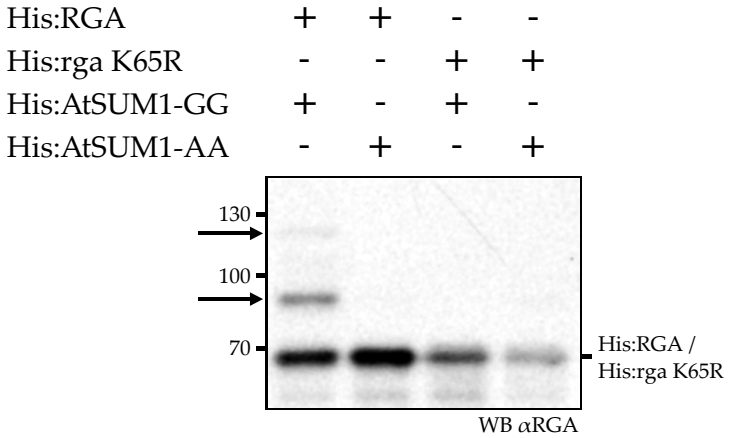


Figure 4.2: Reconstituted SUMOylation assay of RGA. RGA is SUMOylated while the mutagenised rga K65R is not, indicating that lysine 65 is the site of SUMOylation *in vitro*. SUM1-GG is the native active form of SUMO1 while the non-ligating SUM1-AA is used as a negative control for the assay. Arrows indicate SUMOylated RGA, with the higher band corresponding to poly-SUMOylation. All lanes contain total bacterial lysates from each expression reaction. Results published in (Conti *et al.*, 2014). WB: western blot.

While this assay suggests that K65 is a major SUMO site in RGA, the behaviour *in planta* may be different and it is possible that secondary SUMO sites may exist, especially since an E3 is not used in the reconstituted assay. However, there is a high degree of confidence that K65 is the primary SUMO site based on the high levels of SUMOylation at this site in the *in vitro* assay. Further analysis of the

SUMOylation state of RGA in planta should be performed to confirm the results from the reconstituted system using a sensitive technique such as MALDI-TOF mass-spectrometry of purified RGA proteins.

4.3.3 N-terminal fusion tags are not present in recombinant RGA

The original aim of this work was to purify large amounts of SUMOylated RGA and investigate the interaction of the SUMOylated form using a method that would provide quantitative binding data in order to test whether binding kinetics support the model of SUMOylated RGA inhibiting GID1 mediated degradation of the DELLA proteins. In order to achieve this aim, soluble SUMOylated RGA protein needed to be purified. Affinity tag chromatography was chosen for RGA purification as large amounts of protein could be produced. Antibody affinity chromatography was deemed to be prohibitively expensive and only small amounts of protein could be purified due to limited amounts of the bespoke αRGA antibody being available. The reconstituted SUMOylation system used His tagged AtSUM1 and SAE2, so an alternative tag was required for RGA to purify it from the free His:AtSUM1 and His:SAE2. The GST fusion tag was selected as a suitable tag which allowed cost effective purification. RGA was subcloned into pDEST15 to add an N-terminal GST fusion to the protein for use as an affinity tag, however, all attempts to purify the protein using glutathione sepharose beads failed suggesting that expressed protein was either insoluble or there was an issue with the tag.

To investigate this issue, both N-terminal His and GST tagged fusion proteins were expressed as before and then the soluble and insoluble fractions were separated and analysed on a western blot. The blot was probed with αGST and αHis antibodies to show whether the soluble tagged proteins were present. However, the probed western blot showed no immunoreactivity against the expressed proteins (Figure 4.3) at their expected sizes. Control proteins excluded any issues related to the antibodies or western blotting procedure. While the GST fusion showed no immunoreactivity at the expected GST:RGA size, a small fragment corresponding to the size of free GST was detected in the soluble fraction of the bacterial lysate. These data suggested that either there was premature termination of translation or that the protein tag was cleaved from the newly synthesised RGA. Later experiments (Figure 4.4) showed that RGA protein was expressed from the GST:RGA plasmid but the protein only corresponded to the size of RGA and not to the full fusion protein, which excludes premature termination of translation as the tag was at the N-terminal of the protein which would have been synthesised first.

The His fusion, like the GST, showed no reactivity at expected size either, again suggesting that the tag was not present. Whether this tag was cleaved or not could not be discerned from the results as the cleaved His fragment would be too small to detect on the gel used in the assay. Interestingly, comparing the His tagged protein to the GST tagged protein in a later experiment (Figure 4.4), the His tagged version was *larger* than the GST version which is somewhat surprising given that the His tagged protein was expected to be smaller. If both tags were being cleaved from RGA, the smaller size of the

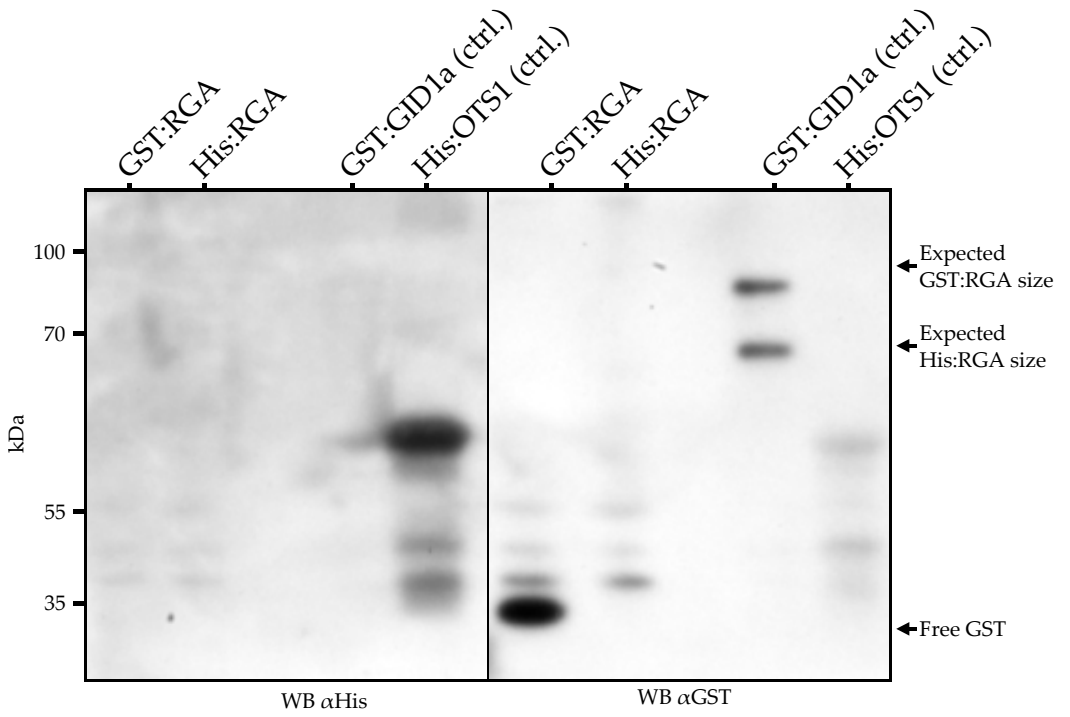


Figure 4.3: No immunoreactivity against N-terminal tagged RGA proteins. GST tagged (pDEST15) and His tagged (pDEST17) RGA N-terminal fusions were probed with αGST and αHis antibodies respectively but no protein of the expected size was detected for either protein. The GST:RGA lane shows a band corresponding to the size of free GST. All lanes, including control lanes, contain total bacterial lysates. Control proteins GID1a and OTS1 were expressed in the vectors pDEST15 and pDEST17 respectively. WB: western blot.

GST:RGA protein suggests that tag cleavage occurs at different sites in the two proteins. However, there is no direct evidence of His tag cleavage in these data.

The fact that purification of GST:RGA failed and that both His and GST tagged RGA could not be detected on a western blot indicated that there was a systematic issue with either the protein itself or the DNA construct that prevented fusion of stable N-terminal tags. Sequencing of the original RGA pENTR plasmid did not show any mutation within the gene or any misalignments with the open reading frame that would have caused issue. The lack of an N-terminal tag meant that purification of RGA, and therefore SUMOylated RGA was not possible and alternative methods for purification were investigated.

4.3.4 SUMOylated RGA produced in *E. coli* is insoluble

In order to purify SUMOylated RGA using the reconstituted *E. coli* system, the SUMOylated form of RGA needed to be in the soluble fraction of the bacterial lysate as functional protein was required for purification and interaction assays with GID1. To test the solubility of the SUMOylated form, both the GST:RGA and His:RGA plasmids were transformed into the reconstituted *E. coli* SUMOylation system as described earlier and the protein fractions from the lysates were analysed by western blot using αRGA antibodies to detect the proteins. The western blot showed that the SUMOylated versions

of both the RGA proteins expressed in the His and GST plasmids were exclusively insoluble (Figure 4.4) while the unmodified RGA protein was present in both fractions. This result was very surprising given that SUMO fusions are widely used to enhance the solubility of difficult to express proteins (Butt *et al.*, 2005; Marblestone *et al.*, 2006; Zuo *et al.*, 2005), though SUMOylation occurs via an isopeptide bond and may behave differently to translational fusions.

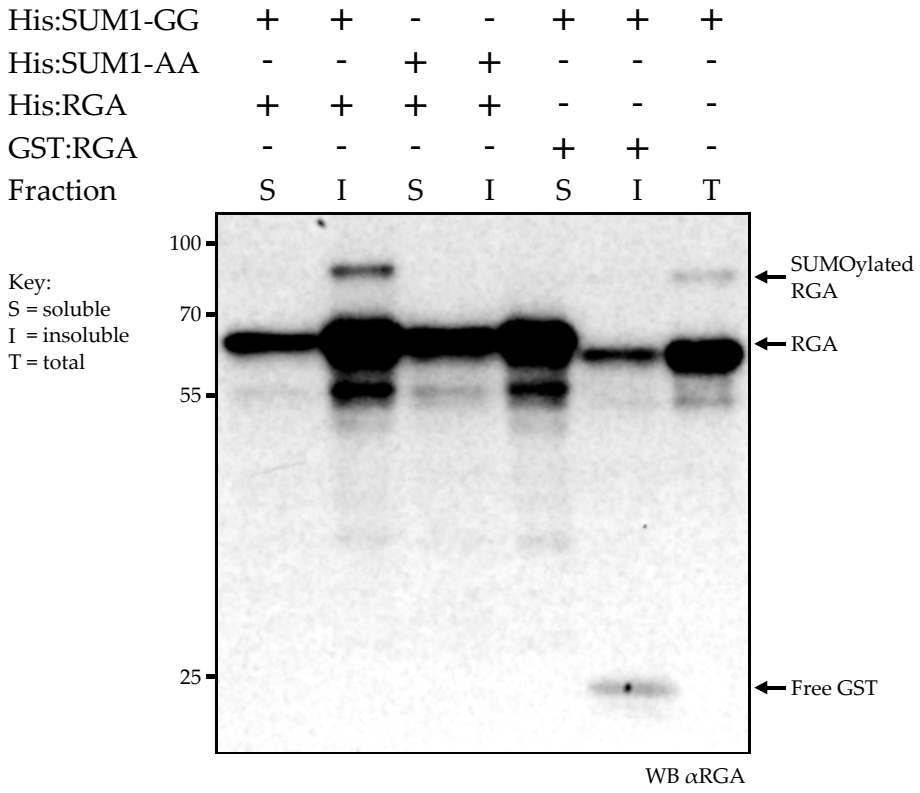


Figure 4.4: SUMOylated RGA produced in the reconstituted *E. coli* system is insoluble. While part of the RGA protein fraction is soluble, SUMOylated RGA is only found in the insoluble fraction of *E. coli* lysate. RGA was expressed in both pDEST17 (N-terminal His tag) and pDEST15 (N-terminal GST tag) but neither improved the solubility of SUMOylated RGA. Furthermore RGA expressed in pDEST15 did not show an increase in molecular weight corresponding to the addition of GST but was in fact slightly smaller the RGA expressed in pDEST17. Additionally a small low molecular weight band is present in the soluble lane, corresponding to the size of free GST. This suggests that the GST portion of the fusion is expressed but a stable GST:RGA fusion is not produced.

The fact that SUMOylated RGA was insoluble meant that unless modifications could be found to the SUMOylation conditions that would render the SUMOylated protein soluble, this system could not be used to produce SUMOylated RGA. N-terminal tag instability added to the difficulties as protein fusions that might improve solubility could not be added at the N-terminus. As an alternative to using the reconstituted system, the use of *in vitro* SUMOylation using purified enzymes could be used but this would again require purified RGA protein. Using *in vitro* SUMOylation also raised the issue of yield as SUMOylation using enzymes is somewhat less efficient. Nevertheless alternative methods of RGA expression in the yeast *Pichia pastoris* (*P. pastoris)* and using C-terminal fusions were investigated which are discussed in the next subsections.

4.3.5 RGA expression in *Pichia pastoris*

Protein misfolding, formation of inclusion bodies (insoluble protein), tag cleavage, premature translation termination, protein degradation and lack of post-translational modification are major issues encountered when trying to express eukaryotic genes in bacteria. The use of eukaryotic systems such as yeast or insect cells can alleviate these issues experienced with bacterial expression (Vincentelli *et al.*, 2005). Due to the issues experienced with RGA tag cleavage in *E. coli* and the large amount of expressed protein forming inclusion bodies, expression in the yeast *Pichia pastoris* (*P. pastoris*) was investigated. *P. pastoris* was selected as it is an easily transformed eukaryotic system which can produce large amounts of recombinant protein with relative ease compared to other eukaryotic expression systems. Expression vectors can be designed to include an α-factor peptide to secrete proteins into the expression media allowing separation of the recombinant protein from most yeast proteins. Using *P. pastoris* to secrete recombinant protein also avoids the issue of yeast cell lysis which is more difficult and less efficient compared to bacteria.

An additional construct was designed to address issues experienced with producing SUMOylated protein. A translational fusion of *AtSUM1* and *rga K65R* was created by a collaborator, *S1:rga K65R*. It was hypothesised that protein produced by this gene would behave in a similar manner to SUMOylated RGA. This construct was included in the *P. pastoris* protein expression experiments.

For the expressed protein to be excreted from *P. pastoris*, an α-factor signal peptide needs to be fused to the recombinant protein which targets the protein to the Golgi apparatus leading to exocytosis. During this process the α-factor is cleaved by endogenous peptidases producing a recombinant protein with minimal additional amino acid sequence. The pGAPZα B plasmid system and the protease deficient strain SMD1168 from Life Technologies were used for expression of RGA and rga K65R. The pGAPZα B plasmid introduces an N-terminal α-factor peptide and a C-terminal His tag to an expressed protein. DNA sequences coding the *RGA, rga K65R* and *S1:rga K65R* were amplified from pENTR plasmids containing these genes and restriction sites were introduced to subclone in-frame inserts into pGAPZα B plasmids using the restriction enzymes XbaI and SfiI (see materials and methods chapter for full details). These plasmids were transformed into *E. coli*. For each construct 5 bacterial colonies were analysed for the presence of the correct insert in the pGAPZα plasmid and all were found to contain the correct size insert (Figure 4.5). Plasmid was extracted from these clones and then sequenced to confirm the correct DNA sequence and subsequently cloned into SMD1168.

For each construct, five yeast colonies were selected and a small scale (10 ml) expression screen was performed to test the expression of the proteins in *P. pastoris*. To test whether His tagged RGA proteins could be purified, 5 ml of the supernatant from the yeast cultures was used in a trial purification using nickel affinity HIS-Select columns (Sigma Aldrich) to isolate His-tagged RGA and rga K65R proteins. Bradford reagent was used to quantify protein from the purification, however, there was no detectable

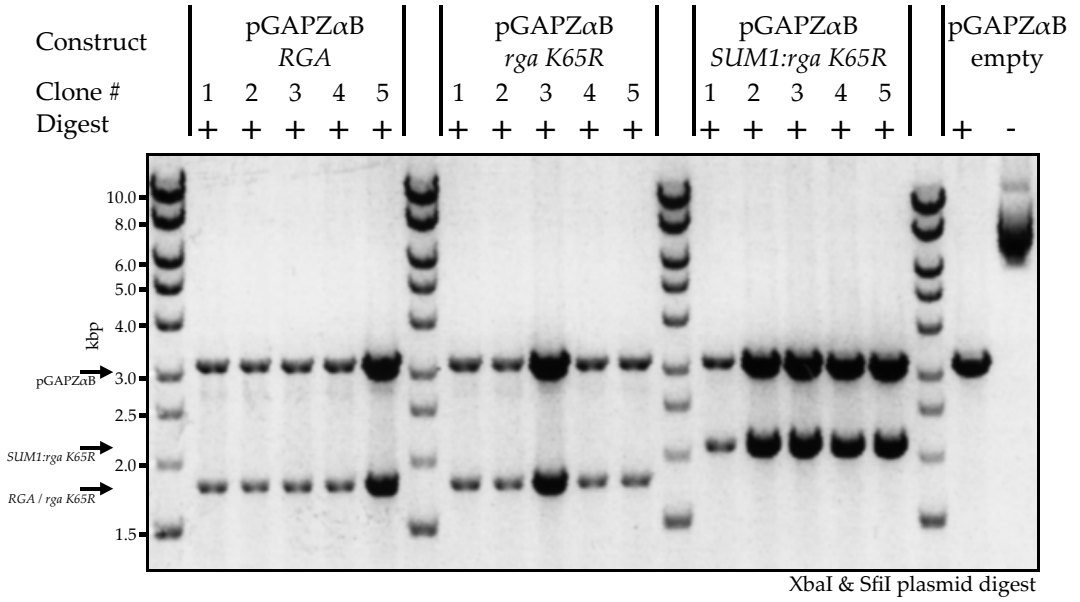


Figure 4.5: Restriction digest of RGA constructs in pGAPZαB to confirm gene insertion. Purified plasmid DNA from 5 independent clones for each construct was digested with XbaI and SfiI to release the cloned gene fragment and confirm insert size. All clones contained the correct insert size and two clones for each construct were selected for sequencing.

protein in any of the eluates. 27 µl of each eluate was then analysed by western blot with αHis antibodies and again no protein was detected, even when the X-ray film was exposed for an extended period of time (Figure 4.6). This trial purification suggested that either no His tagged RGA protein was produced or the protein was expressed but was not excreted into the extracellular media.

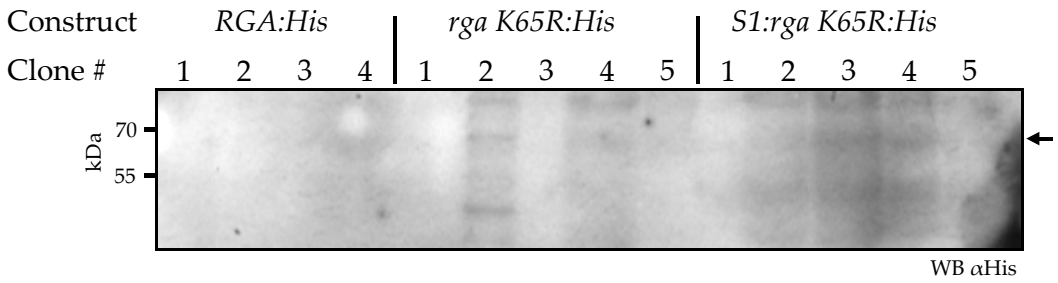


Figure 4.6: Small scale nickel IMAC purification of RGA:His, rga K65R:His and S1:rga K65R:His from *P. pastoris* extracellular media. Arrow indicates expected size of RGA protein. No protein was detected for the RGA:His construct while trace amounts were detected for the other two constructs. Overall no appreciable amounts of protein were purified from the extracellular media. WB: western blot.

To investigate lack of purified RGA from the nickel affinity spin columns, the RGA and rga K65R expression was repeated and both the cellular and extracellular total fractions were analysed using αRGA antibodies to test whether any RGA protein was expressed. The two fractions were treated with a strong denaturing lysis buffer and analysed by western blot and RGA was detected in most of the cellular extracts but not the extracellular extracts. This indicated that secretion did not occur (Figure 4.7). Although two secretion samples showed a band, the corresponding expression cultures evaporated during incubation due to loose fitting covers on the Erlenmeyer flasks used for expression. The samples are

marked with the letter 'b' in Figure 4.7. It is likely that the protein bands seen are due to proteins released from lysed cells in these samples and not from secretion. The RGA proteins observed in the cellular faction migrated at two different molecular weights, both larger than the expected size of 67.4 kDa for RGA after α-factor cleavage. The lower of the two observed bands corresponded to the expected size of RGA fused to the α-factor of 76.8 kDa suggesting that no cleavage of the α-factor peptide occurred. The reason for the presence of the larger band at 100 kDa is not known but may be due to post translational modifications of the protein such as glycosylation as RGA contains a number of predicted glycosylation sites.

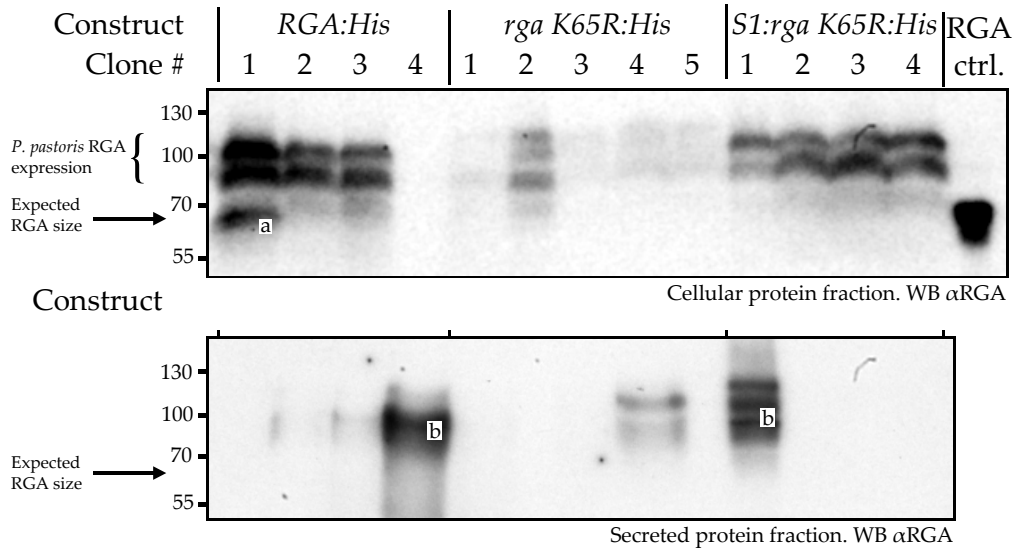


Figure 4.7: RGA expression in *P. pastoris*. RGA constructs in pGAPZαB were expressed in *P. pastoris* and the cellular and secreted fraction was analysed. RGA protein was expressed in the cellular fraction. All cellular RGA proteins migrated with higher molecular weight than the expected RGA size, this discrepancy in size for the lower band is equal to the size of the excretion α-factor fusion. The high band may be due to additional PTMs. The band with the 'a' was due to accidental loading of the RGA control protein with *P. pastoris* lysate in the same lane and should be disregarded. Very little protein was observed in the secreted fractions. Lanes labelled with 'b' showed protein but during expression the yeast media evaporated due to poorly fitting flask covers. The lanes marked with 'b' probably contain contaminant RGA from lysed cells. WB: western blot.

The results of the trial expressions indicated that RGA protein was expressed but was retained within yeast cells. Next a large scale expression of the best expressing yeast clone for the RGA:His was performed (clone 1 for RGA:His) in 500 ml of media and to test whether large scale expression had an affect on secretion, both the cellular and extracellular fractions were analysed. After the yeast culture had grown to saturation, the cells and expression media were separated by centrifugation and the cells were frozen in liquid nitrogen for later processing. The expression media from the expression was then filtered through an 0.4 μm filter and imidazole and NaCl were added to prepare the media for purification on 1ml nickel nitrilotriacetic acid (NTA) His-Trap columns (GE Life Sciences) capable of purifying milligram amounts of protein. The prepared expression media was passed through the His-Trap columns at a rate of 1 ml/min in a cold-room (approximately 8 hours). The column was then

washed and the bound protein was eluted into 1 ml fractions. A Bradford assay was used to test which fractions contained protein, however, no detectable protein was present. Aliquots were taken from the elution fractions expected to contain protein for later analysis.

The frozen yeast pellet was then lysed using the non-denaturing detergent lysis reagent Y-PER Plus (Pierce) and the lysate was clarified by centrifugation and then filtered through a 0.4 μm membrane. After adjusting the media composition by adding imidazole and NaCl, the cellular lysate was then purified using a 1 ml His-Trap column and eluted into 1 ml fractions. Bradford assays showed that a low amount of protein was present in a number of the fractions and these were pooled then concentrated using an Amicon spin column (Millipore) and the concentration of the protein was determined by infrared spectroscopy. A yield of around 50 μg of protein was calculated which was significantly less than the milligram scale yield expected for an expression volume of 500 ml. Background protein was expected in His column purifications from yeast cells as a number of yeast proteins are known to bind to nickel columns. Since the amount of protein purified was so low, it was suspected that the protein present was yeast background protein rather than recombinant protein.

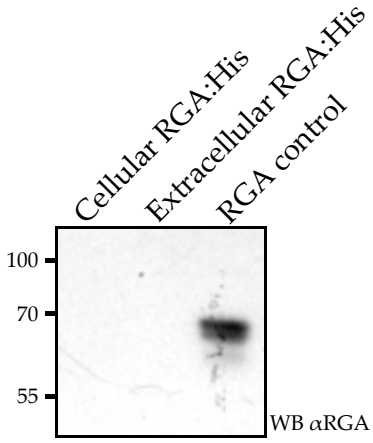


Figure 4.8: Analysis of *P. pastoris* intracellular and extracellular nickel IMAC protein purification. No RGA protein was detected in the elution from the nickel IMAC column suggesting the His tag is either not present or obscured. WB: western blot.

Aliquots of the elution from both the extracellular and cellular fraction were then analysed by western blot using αRGA antibodies. No RGA protein was detectable in either sample indicating that no, or very little, RGA protein was purified (Figure 4.8). The inability to purify RGA or rga K65R protein from either the cellular or extracellular fraction indicated that there was an issue with the C-terminal His tag of the proteins. The issue is not known but the most likely reasons could be tag removal or premature termination of translation, with the latter well documented in yeast expression systems (Henikoff & Cohen, 1984). Additionally, the expressed protein did not appear to be secreted into the extracellular media, this combined with lack of α-factor secretion peptide cleavage suggested the recombinant protein was not correctly processed in the Golgi apparatus within the yeast cells. Interestingly, in yeast two-hybrid experiments performed in *Saccharomyces cerevisiae* (*S. cerevisiae*) as described in Chapter

5, it was noted that yeast cells expressing RGA fused to the GAL4 activation domain (AD) were less viable and if the cells were left on agar plates for two weeks, the RGA containing plasmid was lost. This suggests that RGA may be toxic to the yeast cells. Also, if RGA was fused to the GAL4 DNA binding domain (BD) alongside an unfused AD, the reporter gene for yeast two-hybrid assay was activated, indicating that RGA may be binding to the AD. Given the role of DELLA proteins as transcription factor inhibitors in plants, it is possible that the capacity for RGA to bind to transcription factors extends to yeast proteins and this is responsible for the reduction in viability seen in *S. cerevisiae.* This phenomenon may also be responsible for the lack of protein processing seen in *P. pastoris* caused by RGA binding to and disrupting yeast transcription factors. However, additional experiments would be required to confirm whether this was actually the case.

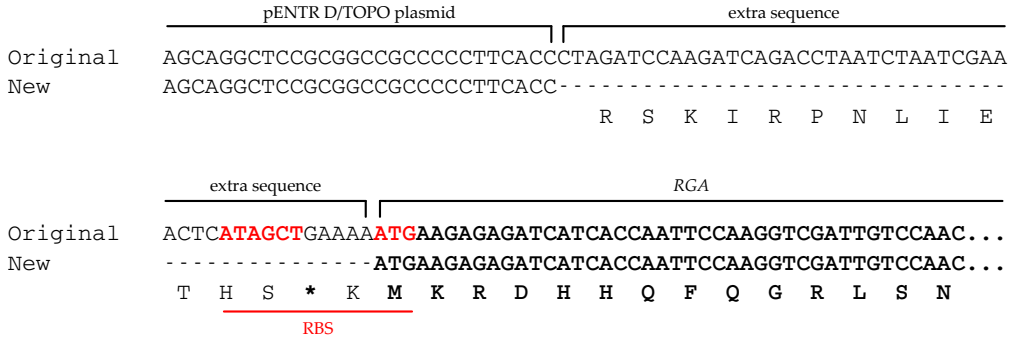
The issue of the lack of protein purified via the His tag could have been further explored to determine whether the His tag was actually present and the pGAPZα B vectors could have been redesigned to remove AT rich regions which are known to interfere with yeast translation of exogenous proteins. However, based on the effort and the additional issue that the α-factor secretion signal peptide was not processed correctly, it was decided that alternative bacterial expression methods would be investigated instead.

4.3.6 Redesigned *E. coli* expression vector for RGA

Attempts to express and purify soluble RGA using affinity tags failed in both *E. coli* and *P. pastoris* expression systems and the *RGA* clone was analysed to identify the cause of the expression issues. Alternative C-terminal fusion tags were also investigated as an alternative strategy to produce tagged RGA that could be purified.

The original *RGA* clone was acquired from a collaborating research group, though the sequence had not been verified independently for our work. The pENTR D/TOPO cloning system uses blunt end cloning and allows insertion of a DNA fragment with the only requirement that the sequence CACC is appended to the 5' end of the sequence to be inserted and therefore allows cloning of DNA sequences with very little additional sequence. The pENTR plasmid is used in the Gateway cloning system to move cloned sequences into pDEST plasmids using homologous recombination to generate tagged expression vectors. The *RGA* clone was expected to only contain the coding sequence for the gene. To verify this the plasmid was sequenced and additional sequence was found upstream of the *RGA* coding sequence which was found to be part of the 5' untranslated region (UTR) from the *RGA* gene. The sequence of the plasmid is shown in Figure 4.9 alongside the expected sequence for blunt end cloning. The additional 5' untranslated sequence introduced a frameshift as it was 47 bases long and also encodes for a stop codon for the sequence in frame with the *RGA* gene showing that any translational fusions introduced by the Gateway system would be out of frame with the N-terminal fusion tags and translation would

5' end of *RGA* gene:



3' end of *RGA* gene:

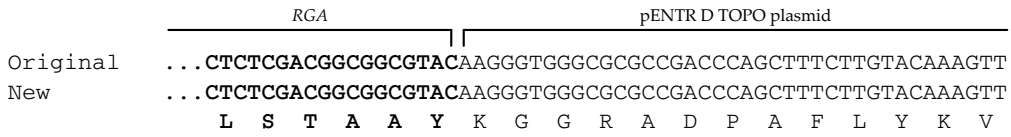


Figure 4.9: DNA sequences of original and new *RGA* pENTR D/TOPO vectors. The original vector contained an extra sequence from the 5' untranslated region of the *RGA* gene while the new vector contained only the *RGA* coding sequence. The additional promotor sequence introduces a frame shift and stop codon into the sequence upstream of the *RGA* gene and also includes a predicted bacterial ribosome binding site which may have been used as an alternative translation initiation site from in old vector.

be terminated before the *RGA* gene. This result was surprising given that this clone had been used to express RGA protein (albeit without an N-terminal tag) which was detected using αRGA antibodies and was of the expected size for RGA. Furthermore, the expression of the tag alone could not account for the size of recombinant protein since the largest tag used was GST with a size of 35 kDa while RGA is 68 kDa.

One hypothesis to explain the observed results was that bacterial translation initiated both at the start of the fusion tag and also at the start of the *RGA* gene from a single mRNA molecule generated by the original *RGA* clone in an expression vector. This would require a ribosome binding site (RBS) upstream of the *RGA* start codon and analysis of the sequence found a putative site, ATAGCT, 6 bp upstream of the *RGA* start codon that conforms to an *E. coli* RBS (Shultzaberger *et al.*, 2001). If such a situation had occurred, a single mRNA generated from an expression plasmid with an N-terminal fusion tag would produce two separate proteins. One protein would be produced for the tag which would be terminated within the extra 5' UTR sequence as this contains a stop codon when in frame with the fusion tag in pDEST plasmids. Another protein would be produced for the *RGA* gene initiated at the alternative RBS. For expression of this clone in pDEST15 which adds a GST, Figure 4.4 shows two proteins corresponding to RGA and GST which is in concordance with this hypothesis. The putative GST band in this figure is results from cross-reactivity to the αRGA antibodies. Figure 4.3 shows that this band is reactive against αGST antibodies.

A new pENTR *RGA* plasmid was created without any additional non-coding sequence, correspond-

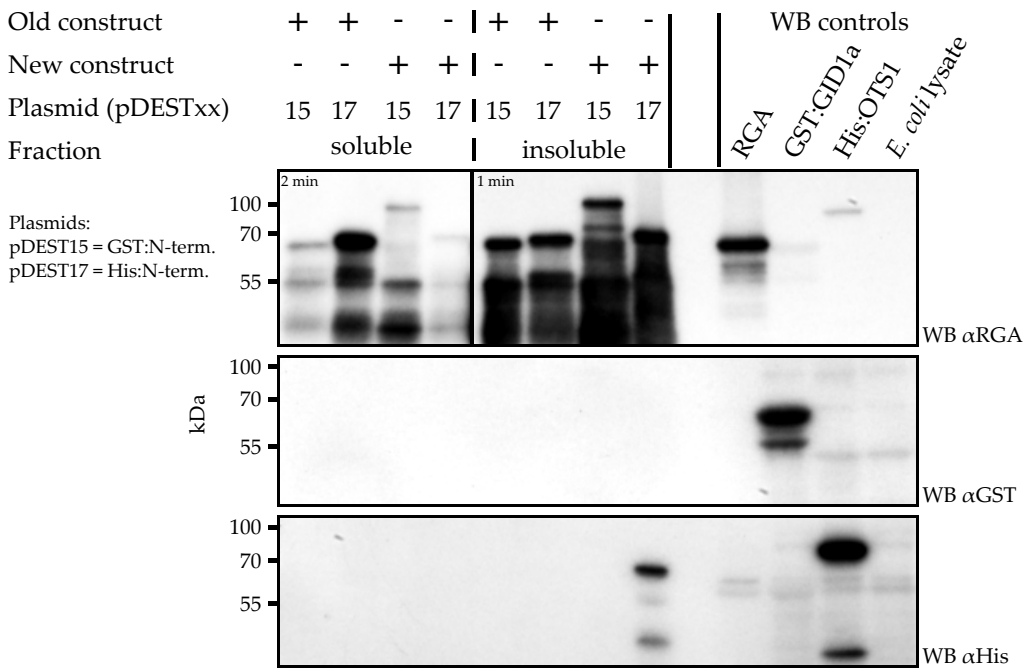


Figure 4.10: Expression of RGA with N-terminal His and GST tags. Two different RGA constructs were inserted into pDEST15 (GST:N) and pDEST17 (His:N) to produce four different expression vectors. None of the vectors produced soluble tagged proteins and only one vector His:RGA using the new construct produced insoluble tagged protein.

ing to the new sequence shown in Figure 4.9 and the sequence was confirmed by DNA sequencing. Unlike the original *RGA* plasmid, expression vectors derived from the new *RGA* clone would be in frame with any tags and would not contain any premature stop codons. The *RGA* gene from the new clone was transferred into pDEST15 and pDEST17 to generate N-terminal GST and His tagged vectors, and with pET DEST 55 to generate a C-terminal His tagged vector (which also introduces an N-terminal StrepII tag, though this tag was not used). The expression of these new vectors was compared to vectors derived from the old *RGA* clone and both the soluble and insoluble fractions were analysed. The results of the N-terminal tagged protein expressions are shown in Figure 4.10 and compared to the old GST tagged vector (pDEST15), the new vector shows the expected increase in protein size. However the protein was not detected by αGST antibodies, the reason for this in not known. In a repeat of the GST:RGA expression using the new clones, GST:RGA was detected with αGST antibodies however, almost all of the protein was insoluble (Figure 4.14). The His tagged protein (pDEST17) on the other hand did show immunoreactivity against αHis antibodies, however, this protein was only present in the insoluble fraction. Nevertheless these results indicated that the new N-terminal expression vector did produce His and GST tagged RGA protein.

The results for the C-terminal His tagged RGA expression vectors are shown Figure 4.11 comparing vectors generated from the old and new *RGA* clones. Both vectors produce His tagged RGA that can be detected with αHis antibodies showing that C-terminal His tags are stable in RGA. Importantly the C-terminal His tagged RGA proteins are present in the soluble fractions of the bacterial lysates. This

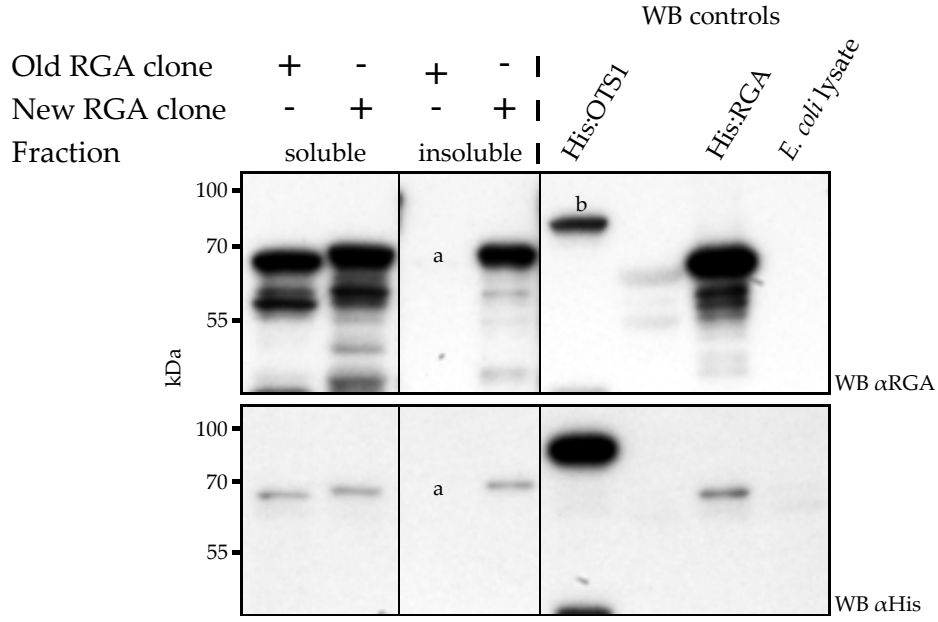


Figure 4.11: Expression of RGA with a C-terminal His tag. The two *RGA* clones were inserted into pET DEST 55 which added an N-terminal strepII tag and a C-terminal His tag. Soluble C-terminal His tagged protein was produced for both *RGA* constructs. (a) Although there is no band for insoluble RGA:His in these lanes, it is probably due to loss of protein pellet during fraction purification. (b) Cross reactivity of the αRGA with the His:OTS1 lane in the His control lane. WB: Western blot.

indicates that soluble tagged RGA protein was produced and that the C-terminal His tagged expression vector derived from the new *RGA* clone could be used to produce purified RGA protein.

The results from this section show that the problems experienced with trying to express tagged RGA in *E. coli* were due a pENTR vector containing an additional sequence that introduced a frameshift mutation and premature stop codons in all reading frames which made N-terminal protein fusions impossible. By removing this extra sequence, N-terminal tags could be added to RGA but no useful soluble protein was produced from these expressions. Ultimately expression vectors derived from pET DEST 55 with a C-terminal fusion tag resulted in production of soluble, His tagged protein which could be purified by IMAC.

4.3.7 Cell free *in vitro* enzymatic SUMOylation of RGA

After a method to produce soluble, tagged RGA had been developed, a method to produce SUMOylated RGA *in vitro* was developed. Since SUMOylated RGA produced in the reconstituted *E. coli* SUMOylation system (Okada *et al.*, 2009) became completely incorporated into inclusion bodies, this method could not be used, instead a free enzyme system was developed using E1, E2 and SUMO proteins based on a modified method by Park-Sarge & Sarge (2009).

Expression vectors for AtSUM1 and an E1 heterodimer consisting of a dual vector containing *SAE1a* and *SAE2* were used from the reconstituted system from Okada *et al.* (2009). The E2 enzyme, *SCE1*, was cloned from *Arabidopsis* cDNA into pDEST17 and the pET DEST 55 vector containing *RGA* described

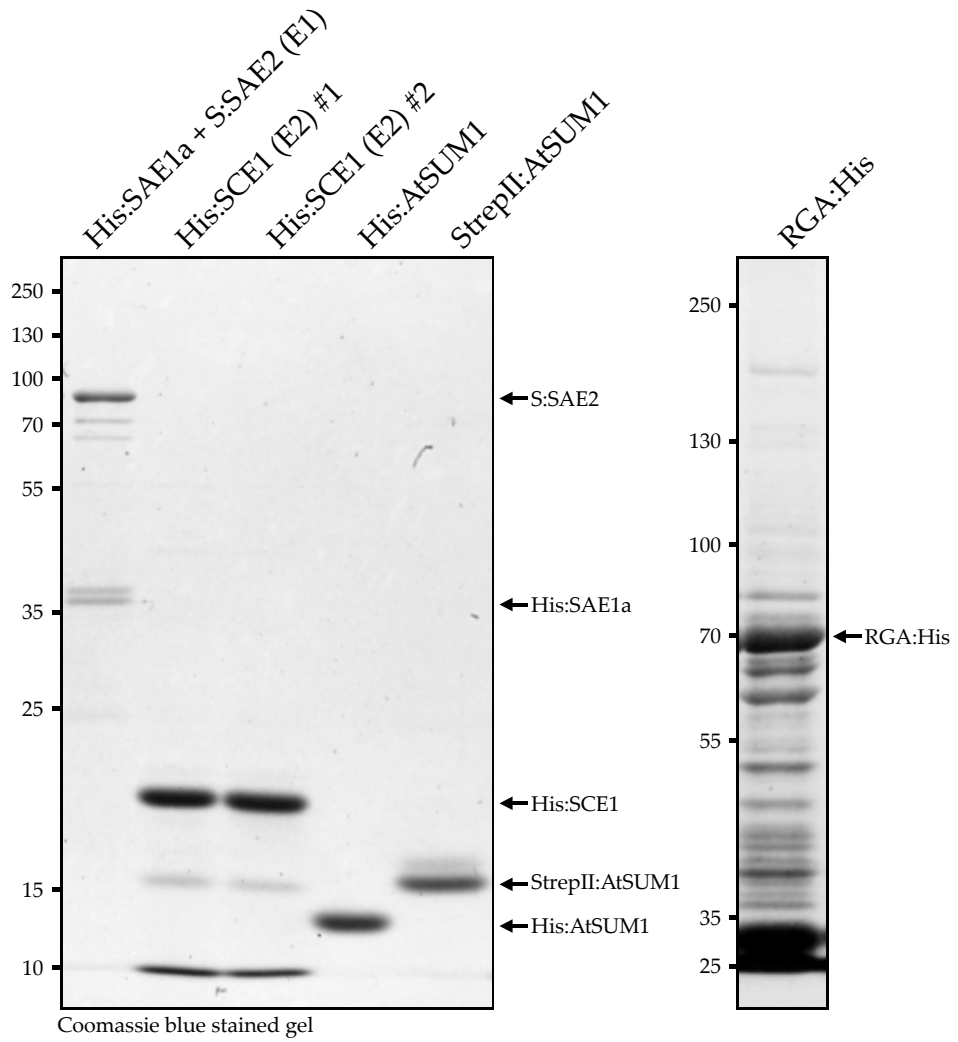


Figure 4.12: Purified proteins for *in vitro* SUMOylation assay. The E1 heterodimer was expressed as a dual insert construct and the full dimer was co-purified via the His tagged SAE1a as the association of the SAE2 protein is strong to maintain the complex during purification. Two different SCE1 clones were expressed and both produced enzyme. His tagged and StrepII SUMO were also purified. RGA:His was purified but showed a large amount of degradation products which was expected based on results from earlier expression tests. The expressed proteins all matched their expected sizes.

earlier was used. All vectors were expressed in BL21 CodonPlus (DE3) RIL *E. coli* cells (Agilent Technologies) and purified using nickel NTA His:Trap columns. His:SEA1a and S:SAE2 were coexpressed, and due to the affinity of the two protein subunits to each other, S:SAE2 copurified with His:SAE1a. All purified proteins were then concentrated and buffer exchanged into a 50 mM Tris buffer at pH 7.6. Aliquots of the purified proteins were analysed by SDS-PAGE (Figure 4.12) which confirmed the expected size of the proteins. The purified RGA:His showed partial degradation of the protein, though a significant proportion was intact. The RGA protein was also found to irreversibly precipitate at high concentrations of the protein ($\gg$1 mg/ml) at low temperatures so the protein concentration was maintained at 1 mg/ml for storage.

For the SUMOylation assay, 20 µl reactions were set up using 50 ng E1, 50 ng E2, 5 µg RGA:His, 5 µg His:SUM1, 1 U pyrophosphatase, 20 mM ATP in reaction buffer (50 mM Tris, 50 mM KCl, 5 mM

MgCl, pH 7.6). Control reactions were set up lacking either ATP, His:SUM1 or RGA:His. The reactions were incubated at 37°C for 1 hour and the reactions were stopped by the addition of 7 µl 4x SDS PAGE sample buffer and were then analysed by western blotting. RGA:His protein was successfully SUMOylated in the assay but only a small amount of protein was SUMOylated which could only be observed using the more sensitive αAtSUM1 antibodies (Figure 4.13). The RGA protein itself showed significant degradation during the assay with only a small pool of protein of the expected size of around 70 kDa remaining after the reaction, this was at least one factor responsible for the small amount of SUMOylated RGA present after the reaction. Nevertheless these data show that SUMOylated RGA can be produced using cell free *in vitro* SUMOylation, however, the method needs to be improved to produce useful amounts of SUMOylated RGA for interaction assays.

Stability of the RGA protein during purification and the SUMOylation reaction was the most significant issue that resulted in low yield. A number of modifications to the methods used could be made to alleviate the stability issue of the protein. Carry over of trace amounts of proteases could contribute to the degradation of RGA and more stringent purification of the RGA protein could reduce protease mediated degradation. Tandem affinity purification using two tags can be used to achieve higher purity as fewer background bacterial proteins are able to bind to the two different affinity media. The pET DEST 55 vector used to express RGA adds both an N-terminal StrepII tag and a C-terminal His tag and tandem affinity purification against these two tags could be used. This method would also remove partial RGA fragments from the expression, as only full length protein would have both tags required for purification. The integrity of the N-terminal StrepII would need to be tested on RGA. This was investigated, however, the StrepII antibodies used were found to non-specifically bind to a bacterial protein of around 70 kDa, the same size as RGA, so the results from this experiment were inconclusive, though the data did suggest that the RGA protein was StrepII tagged as soluble degradation products were seen in the StrepII:RGA sample but not the control (Figure 4.14). However, different αStrepII antibodies, which do not cross-react with bacterial proteins need to be used to confirm whether or not soluble RGA is StrepII tagged.

The rate of SUMOylation in the assay could also be increased by including a SUMO E3 ligase in the reaction mix. At least four SUMO E3 ligases have been identified in *Arabidopsis (Novatchkova et al., 2012)* with SIZ1 and HYP2 the most studied (Ishida *et al.*, 2012) and Miura *et al.* (2009) showed that recombinant SIZ1 protein enhances the rate of SUMOylation of ABI5 *in vitro*, making SIZ1 a good candidate E3 as it has been shown that the functional enzyme can be expressed in *E. coli.*

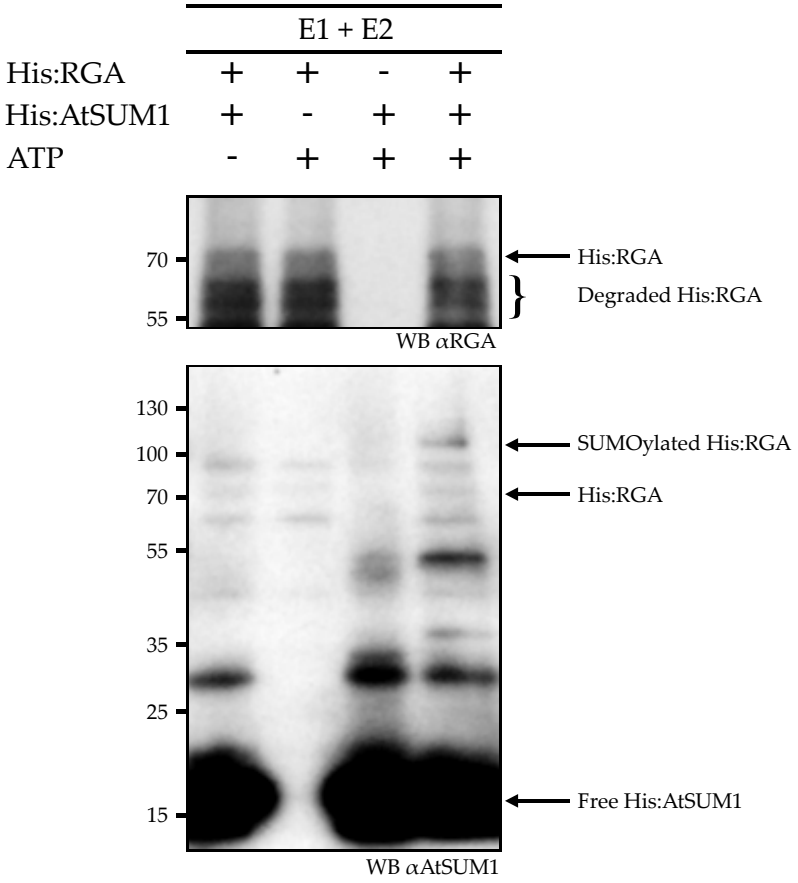


Figure 4.13: *In vitro* enzymatic SUMOylation of RGA. SUMOylation of RGA:His was observed when all AtSUM1 and ATP were present. The amount of SUMOylated RGA produced was low and could only be detected using αAtSUM1 antibodies. Incubation of RGA:His in the SUMOylation reactions lead to a high degree of degradation of the protein which was responsible for the low yield of SUMOylated protein. WB: Western blot.

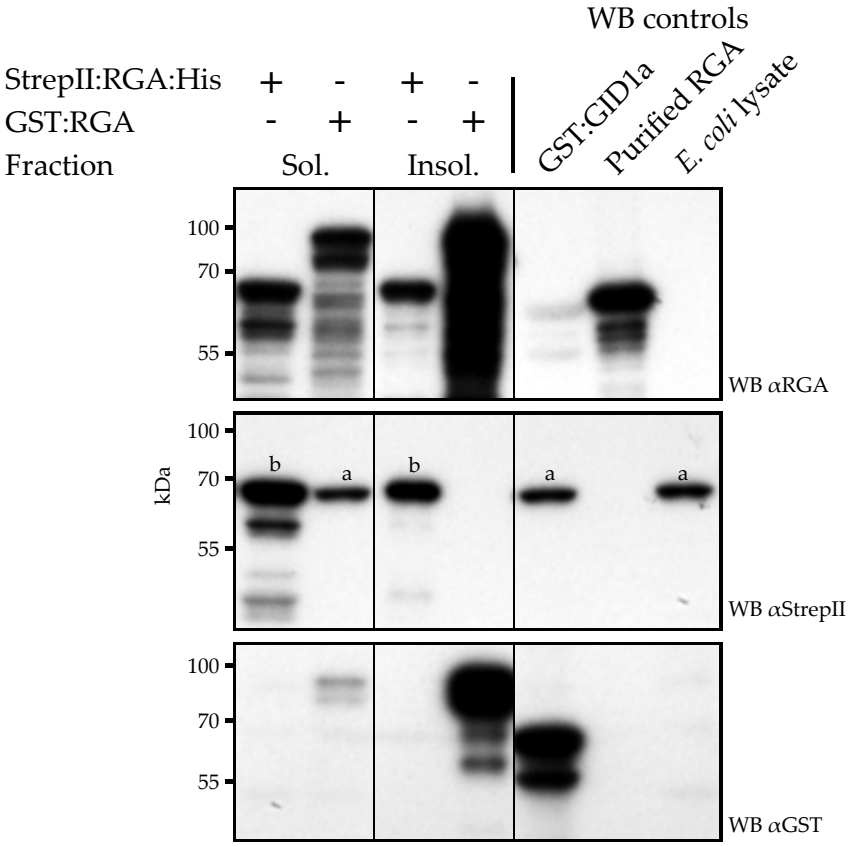


Figure 4.14: Expression of new RGA clone in pET DEST 55 and pDEST15. pET DEST 55 adds an N-terminal StrepII tag and a C-terminal His tag to RGA (StrepII:RGA:His). pDEST15 adds an N-terminal GST tag to RGA (GST:RGA). The αStrepII antibody cross-reacted with a bacterial band at 70 kDa, the same size as StrepII:RGA:His. Due to the cross-reactivity, it could not be demonstrated definitively that RGA with a StrepII was produced, though the band was much stronger for the StrepII:RGA:His lanes (b) than the non specific bands (a). GST tagged RGA was detected, however, almost all of this GST:RGA was insoluble. WB: Western blot.

4.4 Discussion

The work presented in this chapter demonstrates that the DELLA protein RGA is SUMOylated and the site of SUMOylation is K65. The SUMOylated lysine found in RGA is conserved in the other DELLA paralogs in *Arabidopsis* and in orthologs in both eudicot and monocot plant species suggesting a conserved role for SUMOylation in the DELLA proteins. While *Arabidopsis* has five DELLA paralogs, most plant species only have a single DELLA gene. There is a degree of functional diversification in the *Arabidopsis* DELLA proteins which is mostly due to differential expression rather than biochemical differences (Gallego-Bartolomé *et al.*, 2010) suggesting that the function of SUMOylation in the different *Arabidopsis* paralogs should play a similar role. The conservation of the DELLA domain of the DELLA proteins could alternatively be due to the functional role of this domain in binding to the GA receptor GID1. Based on structural evidence (Murase *et al.*, 2008) the SUMO site motif in the DELLA proteins is surface exposed. The high degree of sequence conservation of the SUMO site suggests that this site is the target for SUMOylation in all *Arabidopsis* paralogs and in orthologs in other species. SUMOylation has been demonstrated for *Arabidopsis* RGA and GAI (Nelis, 2011) so far.

The SUMO site in the DELLA proteins in the second α-helix forms one of the binding surfaces that interact with the receptor GID1. Based on the crystal structure of GAI and GID1a binding, the SUMOylated lysine (K48 in GAI) does not participate directly in the interaction with GID1. The location of an attached SUMO moiety at this site could quite reasonably interfere with the DELLA-GID1 interaction by obscuring binding surfaces in the DELLA protein. Interference would need to be tested experimentally to confirm whether this is the case. However, if the model that the SUMO moiety interacts with a SIM in GID1 is correct, a typical interaction assay would not be able to test whether SUMOylation of DELLA proteins blocks their typical binding to GID1. Previous work using Y2H experiments to investigate SUMOylation and SIM binding has found that by mutating amino acids in the binding groove of SUMO, SIM interaction can be abolished (Kroetz & Hochstrasser, 2009). Using a DELLA protein SUMOylated by a non-interacting mutant of SUMO in an interaction assay could overcome the issue of SUMO binding to GID1 and specifically test the DELLA-GID interaction of SUMOylated DELLA proteins.

4.4.1 Expression of RGA

Expression of recombinant DELLA proteins proved to be difficult with issues encountered with introducing fusion tags and with protein solubility. Furthermore SUMOylated RGA produced in a reconstituted SUMOylation system in *E. coli* was completely insoluble. There are very few examples of recombinant DELLA expression in bacteria in the literature. Most work with DELLA proteins used tagged RGA expressed in plants using both stable transgenic lines or transient expression using *Agrobacterium tume-*

faciens to transform plant cells (Gallego-Bartolomé *et al.*, 2012). Studies on the ubiquitin E3 ligase F-box protein SLY1 required for DELLA degradation used recombinant SLY1 to pulldown RGA from plant extracts (Dill *et al.*, 2004) and the original GA dependent binding of the RGA-GID1a interaction studies used recombinant GST-GID1a from *E. coli* to pull down RGA from plant extracts (Griffiths *et al.*, 2006). While the majority of research on the DELLA proteins has used plant protein, His tagged RGA (Tyler *et al.*, 2004) and GAI (Qin *et al.*, 2014) has been reported, however, details of the expression plasmids were not reported so the strategy these authors used to produce tagged DELLA proteins is unknown. Recently Wang *et al.* (2009) successfully expressed RGA in *E. coli* using an N-terminal maltose binding protein (MBP) tag and purified the protein using dextran affinity chromatography. For some difficult to express proteins, the addition of an MBP tag can increase the amount of expressed protein in the soluble fraction (Kapust & Waugh, 1999).

The initial problems encountered with RGA expression in *E. coli* were due to the inclusion of 5' UTR DNA in the vector received from a collaborator which prevented the fusion of N-terminal tags due to the introduction of stop codons in the open reading frame of the expressed gene. After this issue was corrected by constructing a new expression vector, soluble protein with an N-terminal tag could still not be produced. GST fusions were undetectable and His fusions were insoluble. Using a C-terminal His tag was successful and was used to produce and purify soluble RGA. It is unclear whether the issues encountered with N-terminal tags are due the protein or the Gateway expression system used. The Gateway expression system introduces short peptide linkers at the N- and C-terminal regions of the expressed proteins and it is possible that these may be incompatible with the RGA gene. If this was the case, expression using a different plasmid system may alleviate this issue.

Expression in *P. pastoris* was also investigated and it was found that the fused α-secretion factor was not cleaved from the protein and the protein was not secreted as desired into the expression media. Furthermore, the protein could not be purified through nickel IMAC which was likely due to the lack of a His tag on the protein.

4.4.2 Production of recombinant SUMOylated RGA

The reconstituted *E. coli* SUMOylation system (Okada *et al.*, 2009) would have been an ideal system for producing large amounts of SUMOylated RGA protein if it were not for the fact that SUMOylated RGA produced by this system formed inclusion bodies. Other groups have reported using a reconstituted SUMOylation system in *E. coli* to produce SUMOylated protein and separation of the modified and unmodified protein was achieved by using different affinity tags on the SUMO and target proteins (Lens *et al.*, 2011). In an attempt to address the fact that SUMOylated RGA in the *E. coli* system forms inclusion bodies, SUMOylation of RGA using a cell free system was investigated. It was demonstrated that RGA could be SUMOylated using AtSUM1 and purified E1 and E2 enzymes, however, the efficiency

of the reaction appeared to be low, with little SUMOylated RGA produced. Degradation of RGA during the SUMOylation reaction was partially responsible for the low efficiency. Though SUMOylation RGA was produced, the yield of this method would need to be significantly improved to produce large enough amounts of SUMOylated RGA for use in other assays.

The issue of degradation of RGA which may be due to small amounts of proteases remaining in the purified protein samples used in the reaction would need to be addressed in future work. More stringent purification procedures could be used to reduce the levels of protease carryover, though this would probably reduce the protein yield as well. Proteins were purified using nickel IMAC, and using both a higher wash volume and higher concentration of imidazole in the wash buffer than the concentration used (30 mM imidazole) could be used to increase the purification stringency. Alternatively using a different protein tag for RGA at the C-terminus giving better purity could be used such as GST or StrepII tags.

The cell free SUMOylation reaction efficiency could also be improved by including a SUMO E3 ligase to the reaction. Inclusion of an E3 in a reconstituted *E. coli* SUMOylation system has been shown to significantly increase the level of SUMOylated protein produced (Weber *et al.*, 2014). The use of an E3 ligase may also reduce the reaction time which would lead to less degradation of the RGA protein. Improving the purity of the protein and the use of an E3 together may allow useful amounts of SUMOylated RGA to be produced by a cell free system.

Chapter 5

Identification of a SIM in GID1a

5.1 Introduction

The gibberellin (GA) phytohormones are perceived by the soluble receptor GIBBERELLIN INSENSITIVE DWARF1 (GID1) in plants (Ueguchi-Tanaka *et al.*, 2005). GID1 has a high sequence similarity to hormone sensitive lipases (HSLs) which are found across eukaryotic species and play an active role in lipid biochemistry, however, GID1 has lost its catalytic ability and the active site of this protein has been adapted to bind to GA instead. The co-opting of HSLs as receptors for GA appears to have begun with the divergence of the angiosperms, with these plant species showing sensitivity and specificity in response to GA application. The gross structure of GID1 has remained very similar to that of the HSLs, with both containing a substrate binding pocket in the major protein domain with a smaller lid domain enclosing the pocket (Shimada *et al.*, 2008). In the case of GID1, the pocket is responsible for GA binding. While a large number of different GA forms have been found and characterised, GID1 only binds to a small number of bioactive GAs including GA_1, GA_3, GA_4 and GA_7, all of which contain specific functional groups required for interaction. The binding of bioactive GAs to GID1 leads to closing and stabilisation of the lid domain over the hormone bound pocket and this conformational state recognises and is able to bind to the DELLA repressor proteins (Murase *et al.*, 2008).

The recognition of the DELLA proteins by hormone bound GID1 (Ueguchi-Tanaka *et al.*, 2007) targets the DELLA proteins for degradation and is the mechanism by which GA regulates DELLA stability (Fu *et al.*, 2002) and gene expression by releasing numerous transcription factors from inhibition by DELLAs. The N-terminal domain of the DELLA proteins containing a VHYNP and DELLA motif is required for recognition by activated GID1, binding to the closed GID1 lid domain (Sun *et al.*, 2010). The binding of the DELLA proteins to GID1 initiates degradation of the DELLA proteins via the 26S proteasome. A SKP1–CULLIN–F-box (SCF) ubiquitin E3 ligase complex containing the F-box protein SLY1 (Dill *et al.*, 2004) specifically detects the bound GID1-GA-DELLA complex and recruits molecular machinery that polyubiquitinate the DELLA protein which is subsequently degraded (Griffiths *et al.*, 2006; Fu *et al.*, 2002). Although GID1 mediated degradation of DELLA proteins is the major mechanism behind GA signalling, Ariizumi *et al.* (2008) found that the action of GID1 binding to DELLAs alone is sufficient to release downstream transcription factors from repression. Overexpression of GID1 in a *sly1* mutant background, which cannot degrade DELLAs, partially rescues the dwarf phenotype of *sly1*, however, to what extent the effect of GID1 mediated DELLA repression plays a role in GA signalling remains unknown.

The amount of GID1 present in the cell determines the sensitivity of the cell to GA, with higher levels of GID1 degrading DELLA proteins at a higher rate for a given level of GA (Ueguchi-Tanaka *et al.*, 2005). GID1 sits at the centre of a very complex network with many feedback loops that regulate plant growth and development. While DELLA proteins block plant growth responses, they promote the transcription of GID1 and GA biosynthesis genes leading to a negative feedback loop that stabilises

DELLA repression (Middleton *et al.*, 2012). There is also crosstalk between the GA pathway and other hormone pathways including auxin (Roumeliotis *et al.*, 2012), jasmonic acid (Hou *et al.*, 2010) and salicylic acid (Alonso-Ramírez *et al.*, 2009). This complex regulatory network allows the plant to coordinate robust responses to environmental cues. One well studied example is the stress response observed under drought or high salt conditions where DELLA mediated restraint of growth has been shown to improve a plant's overall ability to survive these conditions (Achard *et al.*, 2006). Growth restraint under stress conditions conserves vital resources that prolongs the ability of the plant to survive while stress conditions persist. The circadian cycle also regulates GA signalling by modulating the levels of GID1, promoting GID1 transcription during the night leading to enhanced GA signalling, with the opposite occurring during the day. Regulation of GID1 by the circadian cycle is responsible for daily rhythmic cycles of plant growth (Arana *et al.*, 2011).

While most agriculturally relevant plant species have a single GID1 gene, *Arabidopsis* has three functional paralogs, *GID1a*, *b* and *c*. These paralogs have a high degree of functional overlap and single knockdown mutants show very subtle phenotypes. Double knockdowns on the other other hand show more noticeable phenotypes that are not explained by gene dose effects alone, suggesting a degree of functional specialisation for the *Arabidopsis* GID1 paralogs. The *gid1a gid1c* mutant shows partial dwarfism while *gid1a gid1b* has reproductive organ deformities and lower seed yield suggesting tissue or organ specialisation for the *Arabidopsis* GID1 paralogs. As expected, triple knockdowns of all three *GID1* genes show a severe dwarf phenotype that is GA insensitive (Iuchi *et al.*, 2007). Nakajima *et al.* (2006) investigated the biochemical differences between the *Arabidopsis* GID1 proteins and demonstrated that while all were able to interact with all five of the *Arabidopsis* DELLA proteins, they exhibited different optimal pH values for GA binding and different overall GA binding affinities. GID1b showed both the narrowest pH binding range and the highest binding affinity for GA, ten times that of GID1a or GID1c. Differences in GA binding affinities of the GID1 proteins are partly responsible for functional diversification of the *Arabidopsis* GID1 proteins, with transcriptional regulation also playing a role (Suzuki *et al.*, 2009).

The results of Chapter 4 demonstrated that the DELLA protein RGA is SUMOylated and a model was proposed whereby SUMOylated DELLA proteins inhibit the ability of GID1 proteins to target DELLA proteins for degradation. One possible mechanism by which SUMOylated DELLA proteins could inhibit GID1 is through direct binding to GID1 in an alternative conformation that is not targeted for degradation. Based on the structure of the proteins and the site of SUMOylation in RGA, the position of the attached SUMO moiety or SUMO chain in the DELLA domain of the protein may prevent binding to GID1 in the typical conformation through steric hindrance suggesting that binding would occur with an alternative conformation. Work by other co-authors on the paper submitted on DELLA SUMOylation demonstrated binding of SUMOylated RGA to GID1a *in vitro* and that this interaction

was GA independent. SUMOylated RGA for these experiment was purified from plant cell lysates (Conti *et al.*, 2014). This is in contrast to the interaction with non-modified RGA which requires the presence of GA (Griffiths *et al.*, 2006). While these results do not show that SUMOylated DELLA binding to GID1 occurs *in vivo*, they do support the proposed model. Since GID1a is able to bind to SUMOylated DELLA in a GA independent manner it was hypothesised that this protein could contain a SUMO interacting motif (SIM) which would be responsible for the observed binding. This chapter describes work that identified and characterised a SIM region in the *Arabidopsis* GID1a protein and investigated the effect of overexpressing versions of GID1a with reduced AtSUM1 affinity in order to test predictions from the proposed model of SUMOylated DELLA inhibition of GID1.

5.2 Chapter aims

- Show that AtSUM1 can interact with GID1a.
- Identify the location of the SUMO interacting motif in GID1a.
- Generate a GID1a mutant that cannot bind to SUMO but retains GA dependant DELLA binding.
- Characterise the phenotype of overexpressing a GID1a SIM mutant in *Arabidopsis.*

5.3 Results

5.3.1 SUMO interacts with GID1a

For the proposed model of SUMOylated DELLA proteins inhibiting GID1, it was hypothesised that GID1 could bind to the attached SUMO moiety or SUMO chain in SUMOylated DELLA proteins. To test whether GID1a could bind to AtSUM1, a co-immunoprecipitation (co-IP) using GST:GID1a and His:AtSUM1 was performed. To test whether the phytohormone GA plays a role in the interaction, an additional co-IP with 10 μM GA_3 in all assay solutions was included. The co-IP (Figure 5.1) demonstrated that His:AtSUM1 interacted with GST:GID1a but not GST alone. AtSUM1 interacted with GST:GID1a in both the presence and absence of GA_3, demonstrating that GA_3 is not required for the interaction. These results suggest that SUMO plays a hormone independent role in GID1 regulation. Although there was a slight difference in the intensity of the AtSUM1 bands in the co-IP result with and without GA_3, it is possible that this difference is due to variability between the different co-IP assays. Quantitative measurements of the interaction would be required to determine whether or not the presence of GA had an effect on the strength of interaction with GID1a.

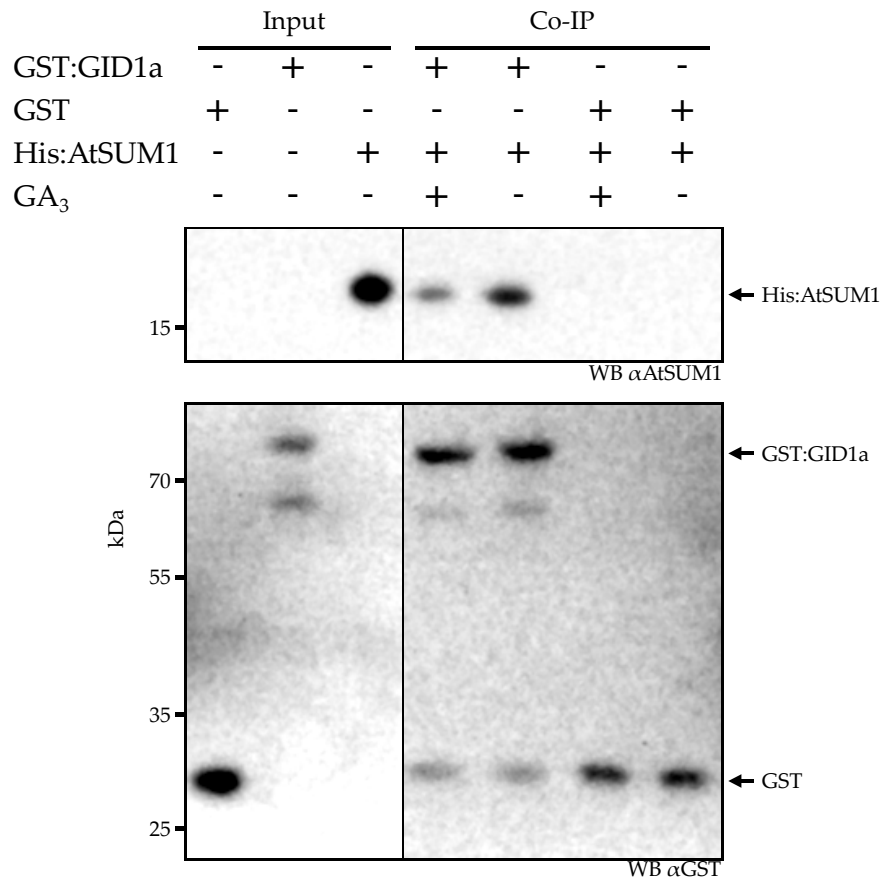


Figure 5.1: Co-IP of His:AtSUM1 with GST:GID1a. His:AtSUM1 was found to interact with GST:GID1a both in the presence and absence of the phytohormone GA_3 demonstrating that the interaction is GA independent. GST:GID1a was immobilised to αGST antibody coated paramagnetic beads and His:AtSUM1 was used as the bait. GST was used as a negative control for the assay and shows that the interaction of His:AtSUM1 was specific to the GID1a component of the GST:GID1a fusion. Results published in Conti *et al.* (2014). WB: western blot.

5.3.2 Bioinformatic analysis of GID1 proteins

Since GID1a had been shown to interact with AtSUM1, it was expected that GID1a would contain one or more SIMs which would be responsible for the affinity of the protein to AtSUM1. The primary and tertiary structure of GID1a were investigated to identify putative SIM sites in the protein that could be tested and mutagenised to generate SIM deficient mutants of GID1a.

The major defining feature of SIMs is their hydrophobic core enriched for leucine, isoleucine and valine. Depending on the SIM orientation, this core has the motif structure of either ΨΨxΨ or ΨxΨΨ which is flanked by either charged or polar residues (Song *et al.*, 2005). To identify putative SIMs in GID1a, regions matching the hydrophobic core motif were identified. The primary sequence of GID1a was screened for sequences consisting of the motif ΨΨxΨ or ΨxΨΨ and 19 such subsequences were found. Two of the subsequences overlapped and were removed since the hydrophobic region would have been too long for this sequence to be a SIM, which left 17 putative subsequences.

For SIMs to be functional they need to be surface exposed and part of unstructured, flexible regions

of a protein. To further eliminate subsequences, the location of the putative SIMs in the 3D structure of GID1a was analysed using the structure resolved by Murase *et al.* (2008) of GID1a bound to GA_3 and the GAI DELLA domain. The majority of the putative SIM-like sequences were found to be within the core of the protein or within secondary structures. All of these sequences were ruled out as SIMs leaving just three SIMs that were surface exposed. The hydrophobic core of SIMs is generally enriched for the amino acids leucine, isoleucine or valine and of the remaining putative SIMs had enrichment of these amino acids. The conservation of these remaining sequences was investigated using all three *Arabidopsis* GID1 paralogs (GID1a, GID1b and GID1c) and homologs from rice, wheat and maize. Two of the remaining putative SIM sequences were found to lie within a highly conserved region of the proteins while the hydrophobic sequence of the third was not conserved in the monocot species. The two putative SIMs in the conserved regions were taken as the best candidates to investigate further. These putative SIMs were named SIM A and SIM B and the cores of the sequences were at positions 15V-18L and 21W-24I respectively.

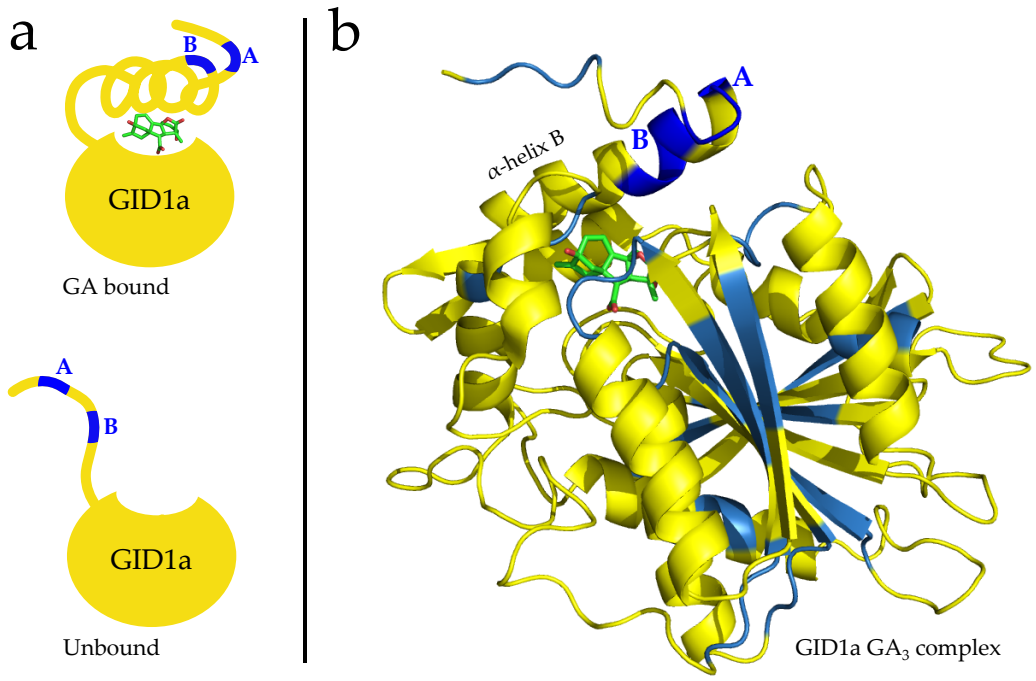


Figure 5.2: Mapping of SIM-like hydrophobic cores onto the 3D structure of GID1a. **a)** Diagrammatic representation of the two conformational states of GID1a based on GA stabilising model showing location of the two putative SIMs. **b)** 3D structure of GID1a with hydrophobic SIM-like patches shown in blue. Most of these regions were within the core of the protein and were excluded (light blue) while the best two candidate SIMs were in the lid domain of the protein and are shown in dark blue. GID1a structure from Murase *et al.* (2008).

The two identified SIM-like sequences lie next to each other in the lid domain of GID1a and these are shown in blue in Figure 5.2. The putative SIM B is within α-helix B of GID1a which is in complex with GA_3. Using molecular simulations Hao *et al.* (2013) demonstrated that the lid domain, including α-helix B, has a high degree of conformational flexibility when GID1 is in the unbound state. Although

this SIM B was within a secondary structure, this putative SIM was retained due to the conformational flexibility of this domain. Due to the conformational flexibility it was speculated that this region might be flexible enough to form an interaction with SUMO.

5.3.3 GID1a SIM peptides bind to AtSUM1

Once the two strong candidate SIM regions in GID1a were identified, oligopeptides of these sequences were tested for interaction with AtSUM1 using far-western blotting. Based on a similar experiment by Namanja *et al.* (2012), 13-mer peptides were designed consisting of the 4-mer hydrophobic core plus six amino acids downstream and three amino acids upstream of this core as the sequences were thought to be reverse type SIMs similar to those investigated by Namanja *et al.* (2012). Namanja *et al.* (2012) demonstrated that SIM interaction with HsSUM1 and 3 could be abolished by disrupting the SIM core region by substituting one of the large hydrophobic residues with a small hydrophobic amino acid such alanine. Based on these data, mutagenised versions of the two SIMs were generated by replacing the second amino acid within the hydrophobic SIM cores with alanine with the intention of creating mutant peptides which do not interact with AtSUM1.

High purity samples of these peptides were then synthesised by Cambridge Research Biochemicals and the peptides were purified by high performance liquid chromatography (HPLC) which was also used to demonstrate sample homogeneity. 1 μg of each peptide was then spotted onto a nitrocellulose membrane and then tested for interaction with AtSUM1 by far-western blotting (Figure 5.3b). The results are shown in Figure 5.3a and demonstrate that the first peptide (SIM A) did not interact but the second (SIM B) did and the mutagenised version (SIM B V8A or V22A in full length protein) showed weaker interaction rather than abolishing the interaction altogether. This far-western blot was repeated twice more with similar results, with SIM B V8A consistently showing weaker interaction. To confirm that SIM B was functionally conserved, the corresponding peptides in a number of GID1 homologues were synthesised, this time using using the SPOT peptide array synthesis method (see Materials and Methods section; the peptide array was synthesised by a collaborating group at Glasgow University and not by myself). The SIM B region of GID1a was highly conserved, with very few differences between the evolutionarily distant monocot species; the far-western blotting assay of these peptides showed that all of the peptides interacted with AtSUM1 (Figure 5.3c) suggesting that SUMO binding capacity of this region of GID1 was conserved and that it may serve an important biochemical role.

Based on previous research by Namanja *et al.* (2012), the V8A mutation to SIM B was expected to completely abolish AtSUM1 interaction and initially the reason for the partial interaction was unknown. Later, as progress was made on the large scale SIM peptide screen detailed in Chapter 3 and AtSUM1 SIM sequences were better characterised, it was realised that the functionally important core sequence may not have been the residues originally identified (amino acids 7-10 in peptide) but rather hydrophobic

residues towards the N-terminal of the SIM B (amino acids 1-4) and that the SIM peptide conformed to SIM type A rather than type R. The random forest SIM predictor described in Chapter 3 supported this notion with a high confidence (predicted FPR = 0.08) that the conformed to type A. For type A SIMs, amino acids after the carboxy end of the hydrophobic core are important in establishing polar interactions or charged interaction with SUMO and the V8A mutation was in this region in the updated model of SIM B. The fact that the V8A mutation does not abolish AtSUM1 interaction supports the notion that this amino acid is not part of the SIM core but rather upstream of it and the reduction in interaction can be attributed to interference of a non-essential region of the SIM. However, this re-evaluation of the residues that constitute the SIM core do not invalidate the interaction results but rather it was a refinement of the GID1a SIM model and the V8A weakly interacting mutant was used to investigate the effect of a weak SIM *in planta*.

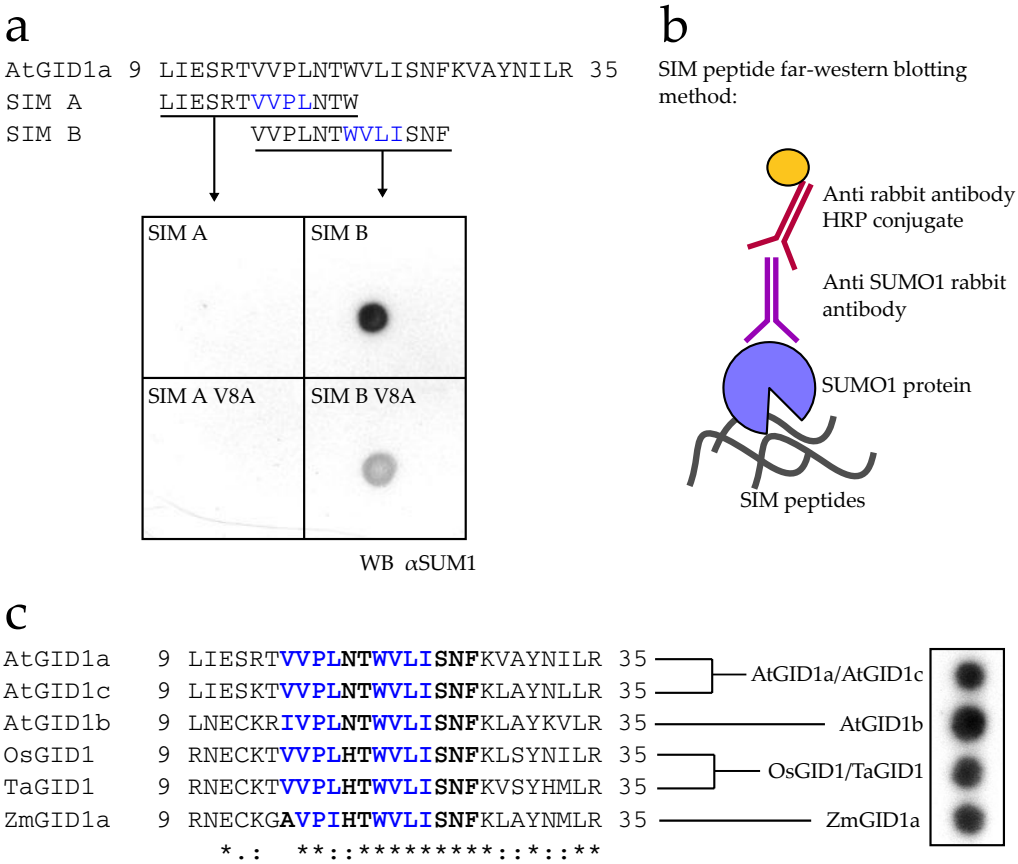


Figure 5.3: Interaction of GID1 SIMs with SUMO1. **(a)** Far-western blot of AtSUM1 against two putative SIMs and mutagenised versions of these peptides. SIM B was found to interact with AtSUM1 and the V8A mutation reduced this interaction. **(b)** Far-western blotting method. AtSUM1 was used as a probe and two antibodies were then used to detect any interaction of the probe protein with peptides immobilised on a support matrix. **(c)** Far-western blot of SIM B from various GID1 homologs all showing interaction with AtSUM1 indicating that the SUMO binding capacity of this region of GID1 is conserved. Species used: At = *Arabidopsis thaliana*, Os = *Oryza sativa* (rice), Ta = *Triticum aestivum* (wheat) and Zm = *Zea mays* (maize).

5.3.4 GID1a SIM mutants maintain receptor function

The mutant SIM B V8A peptide had been shown to be a weaker AtSUM1 interactor and the mutation was introduced into a clone of the full GID1a coding sequence generating the mutant gid1a V22A. V22 lies within a highly conserved region of the GID1a lid component and alanine scanning analysis of the rice GID1 homolog by Ueguchi-Tanaka *et al.* (2007) has shown that disruption of this region can abolish the GA dependant interaction with the DELLA proteins; a triple alanine substitution of residues 20T - 22V was found to abolish interaction with the rice DELLA protein SLR. Additionally Yamamoto *et al.* (2010) showed that the P99S mutation in the rice GID1 leads to GA independent binding of the receptor to DELLA proteins. The binding of the mutant gid1a V22A to RGA was investigated with a yeast two-hybrid (Y2H) assay to test whether DELLA binding was abolished and whether the receptor maintains GA dependent binding. Another mutant, gid1a V22S which was generated as an alternative to the V22A mutation was also tested. The mutant gid1a V22S was predicted to also have weaker interaction with SUMO.

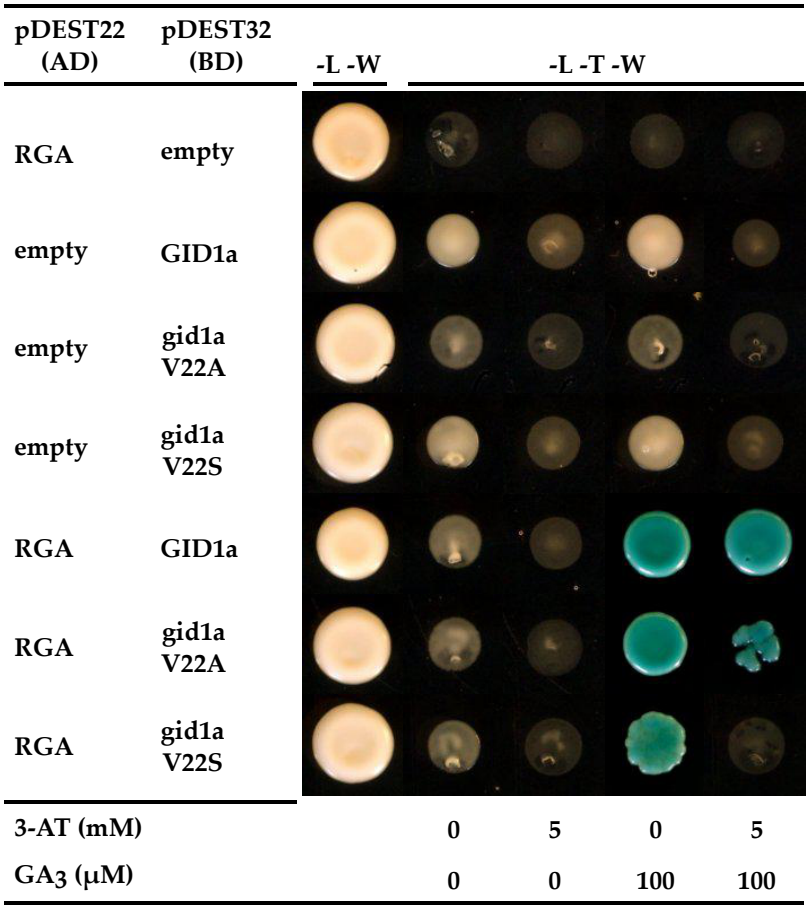


Figure 5.4: The SIM mutants gid1a V22A and gid1a V22S maintain GA dependant binding to RGA. All three proteins only interacted with RGA in the presence of GA_3. The *HIS3* reporter gene inhibitor 3-AT was used to increase the interaction assay stringency. All constructs were grown on minimal nutrient agar lacking leucine and tryptophan (-L -W) media to select for the two plasmids containing each gene. Interaction was shown by growing the yeast on media lacking leucine, tryptophan and histidine (-L -W -H) supplemented with X-α-gal which turns the yeast colonies blue upon activation of the *MEL1* galactosidase reporter gene.

The Y2H assay included yeast clones grown on agar plates with and without GA_3 to test GA dependance. The wild type protein showed the expected GA dependant interaction with RGA which was not affected by the addition of the reporter gene inhibitor 3-Amino-1,2,4-triazole (3-AT) (Figure 5.4). The two GID1 mutants also interacted with RGA and exhibited GA dependance in the interaction, however, addition of 3AT resulted in a slight reduction in the colony size of gid1a V22A and very large reduction for gid1a V22S. The Y2H assay was repeated twice more using yeast cells from different transformation events and the assays gave similar results. Although Y2H experiments are semiquantitative, the results from these assays suggest that although gid1a V22A interacts with RGA in a GA dependant manner, the interaction is weaker than the wild type and the interaction with the gid1a V22S mutant is even weaker still. Since gid1a V22A retained GA dependant binding to the DELLA protein RGA and the interaction was stronger than gid1a V22S, this mutant was used for further investigation *in planta.*

5.3.5 Investigating GID1a interactions using SPR

To investigate the binding strength of AtSUM1 and RGA to GID1a and the two mutant forms of the protein, surface plasmon resonance (SPR) was used. The GID1a sequences in the vector pENTR D/TOPO were subcloned into the plasmid pDEST15 to generate N-terminal GST fusion clones and the proteins were expressed in CodonPlus RIL (DE3) *E. coli* cells and purified using a glutathione sepharose column. Figure 5.5 shows an SDS-PAGE gel of the purified recombinant GID1a proteins. His:AtSUM1 and RGA:His protein from work described in Chapter 4 was used (Figure 4.12). A major problem encountered with the GID1a proteins was degradation of the protein which liberated a small free GST fragment (confirmed by western blot analysis with αGST antibodies). To purify sufficient amounts of protein, the purification procedure time had to be kept to a minimum and all purification steps had to be performed on ice to keep the protein and solutions close to 0°C.

For the SPR assay, the proteins were covalently coupled to a CM5 SPR sensor chip (GE Life Sciences). All reactions were carried out in a 20 mM phosphate buffer pH 7.4 with 150 mM NaCl. In order to bind the proteins to the chip, the pH of the protein buffer solution was optimised to attract the proteins to the chip surface through electrostatic attraction. The GID1a and mutant gid1a protein solutions appeared to have a biphasic response with peaks at pH 4.5 and 5.0 compared with the GST control protein which had a single peak at pH 4.5. The predicted isoelectric point for GID1a is 6.6 and for GST is 4-5. Therefore for the GID1a solutions it was hypothesised that the lower value pH peak corresponded to free GST while the higher to GST:GID1a or GID1a protein and the higher pH values for these proteins were used with the coupling reaction. Despite scouting for optimal pH values, the capture efficiency of GID1a and the mutant proteins was very low and below the target binding of 2000 response units (RU). The final covalent protein binding results are shown in table 5.1.

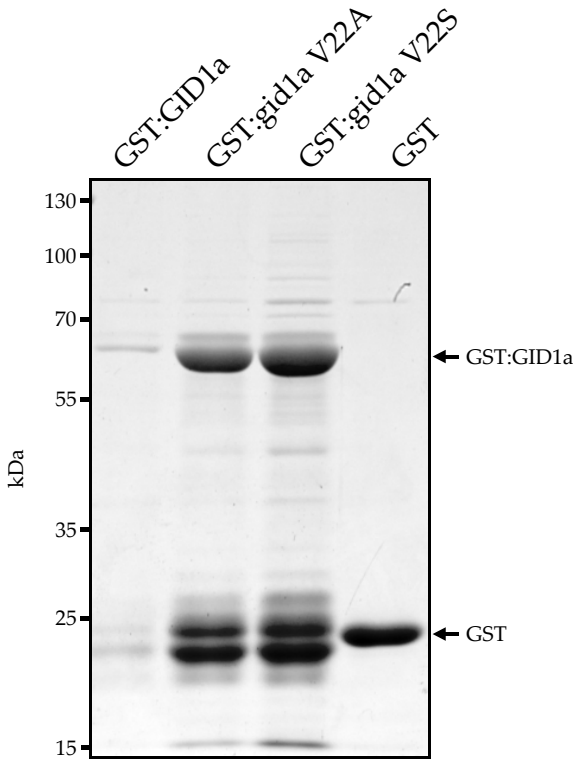


Figure 5.5: GID1a mutant protein purification. GST:GID1a, GST:gid1a V22A, GST:gid1a V22S and GST proteins were purified using a GSTrap glutathione affinity column and moderate levels of GST:GID1a degradation products are present. The band for the wild type GID1a protein is weaker as less protein was loaded onto the gel in error due to different concentration of this sample.

The poor coupling observed for GID1a and the mutant proteins could have been due to the complex mixture of intact proteins and degradation products. Ideally the GST should have had an RU of 40% of the GST:GID1a proteins, as this corresponded to the proportion of GST by mass in the GST:GID1a fusion proteins. The GST control channel was used as a control and subtracted from the GST:GID1a protein channels. Subtracting this channel from the others removed the interaction signal from non-specific interactions with the GST protein and from the CM5 chip surface. The significantly higher level of bound GST protein compared to GST:GID1a protein compromised the accurate removal of non-specific interaction signals. Nevertheless the interaction with the proteins in table 5.2 was investigated to test whether GA dependance with RGA could be observed and whether AtSUM1 could interfere with the GID1a-RGA interaction. The response values for each protein were normalised by the total amount of immobilised protein bound and then the GST channel was subtracted from the three GID1a protein channels. These results were then normalised by the original amount of GID1a immobilised.

Ligand	pH	Final protein response (RU)
GID1a	4.8	376.4
gid1a V22A	5.2	433.9
gid1a V22S	5.2	316.6
GST	4.5	2314.9

Table 5.1: GID1a mutant protein coupling to the CM5 chip for SPR analysis. The coupling of GID1a, gid1a V22A and gid1a V22S was poor and below the target of 2000 RU while the coupling of GST was satisfactory.

After normalisation, apparent binding signals could be seen for all ligands tested and these are shown in Figure 5.6. However, all of the ligand binding treatments show the same trend, with GID1a and gid1a V22S showing the same response and gid1a V22A showing a weaker response. The addition of GA_3 to the samples had no effect on the responses and RGA in the absence of GA_3 still produced the same strength signal. These observations strongly suggested that the response seen in this SPR assay was not from ligand binding to the various GID1a proteins but rather non-specific binding. Due to the difficulties encountered with immobilising the GID1a proteins to the CM5 chip, it is possible that little or no GID1 protein was immobilised to the chip. The response signal seen is most likely due to mismatched non-specific interaction subtraction resulting from incorrect proportions of GST to GST:GID1a bound to the CM5 chip.

Run #	Protein(s)	100 μM GA_3
1	AtSUM1	no
2	AtSUM1	yes
3	RGA	no
4	RGA	yes
5	RGA + AtSUM1	no
6	RGA + AtSUM1	yes

Table 5.2: Interactors tested against GID1a and GID1a mutant proteins. Each assay condition was repeated once with similar response results.

Should this experiment be repeated in the future, purification of a homogenous GID1a sample could alleviate the issues encountered. The degradation of the GST:GID1a fusion which decouples the tag from the fusion protein was the greatest issue encountered with the stability of this protein. Using a specific protease cleavage site to remove the fusion tag during purification would allow the preparation of GID1a without the GST tag which may improve the stability of the protein and would reduce the number of protein species in the sample. Additionally size exclusion chromatography could be used to

separate full length proteins from degraded protein fragments allowing the purification of a homogenous protein sample.

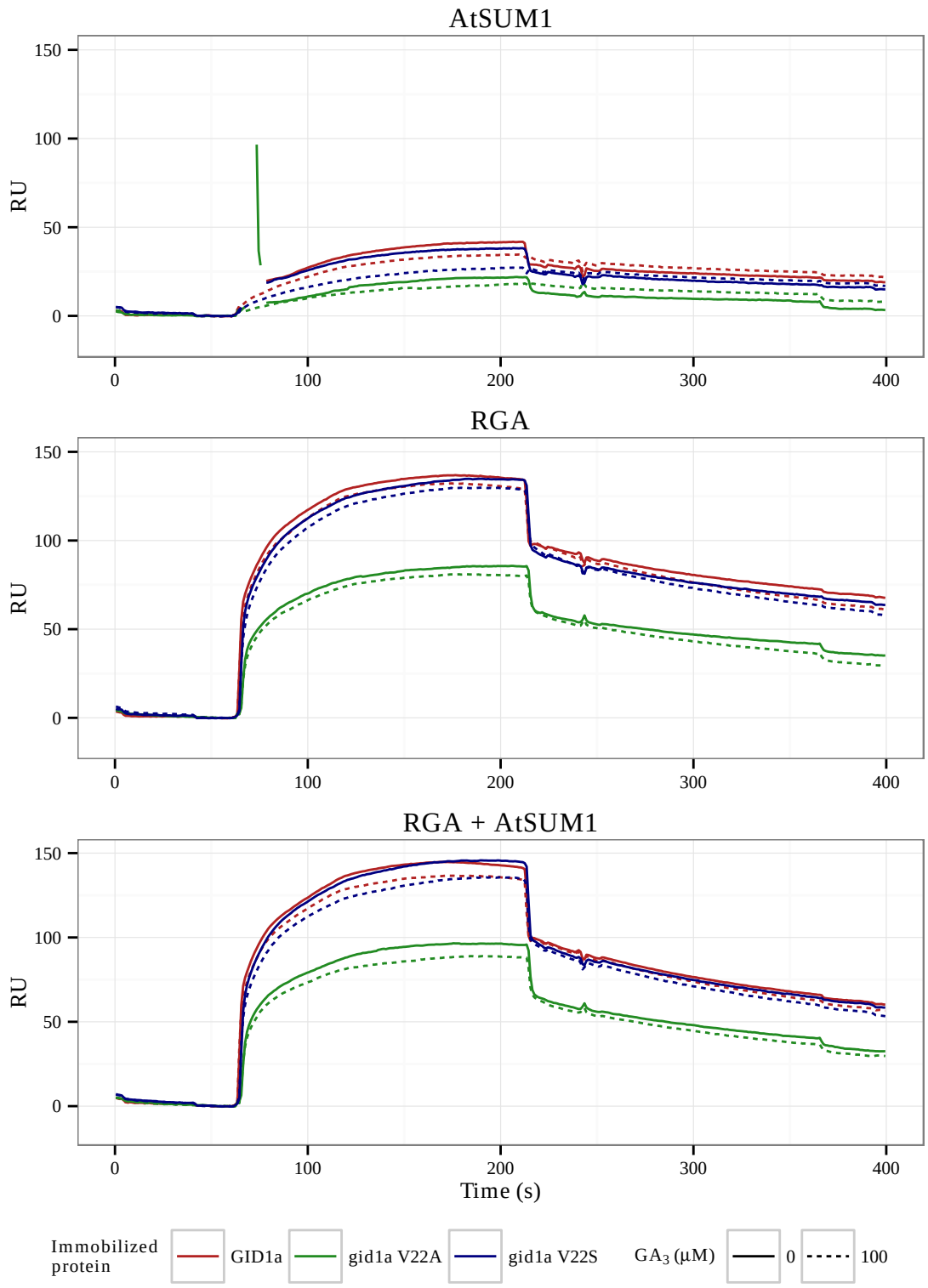


Figure 5.6: SPR sensograms of the binding between the various GID1a proteins and AtSUM1 and RGA. The expected molecular interactions were not observed including GA independent binding of RGA. Based on these observations and from the number of technical difficulties experienced with this assay, it is likely the interaction results observed in these sensograms are due to non-specific interactions. No conclusions were drawn from these data regarding the interactions of the proteins investigated.

5.3.6 Exposure of *Arabidopsis* to a synthetic SIM peptide

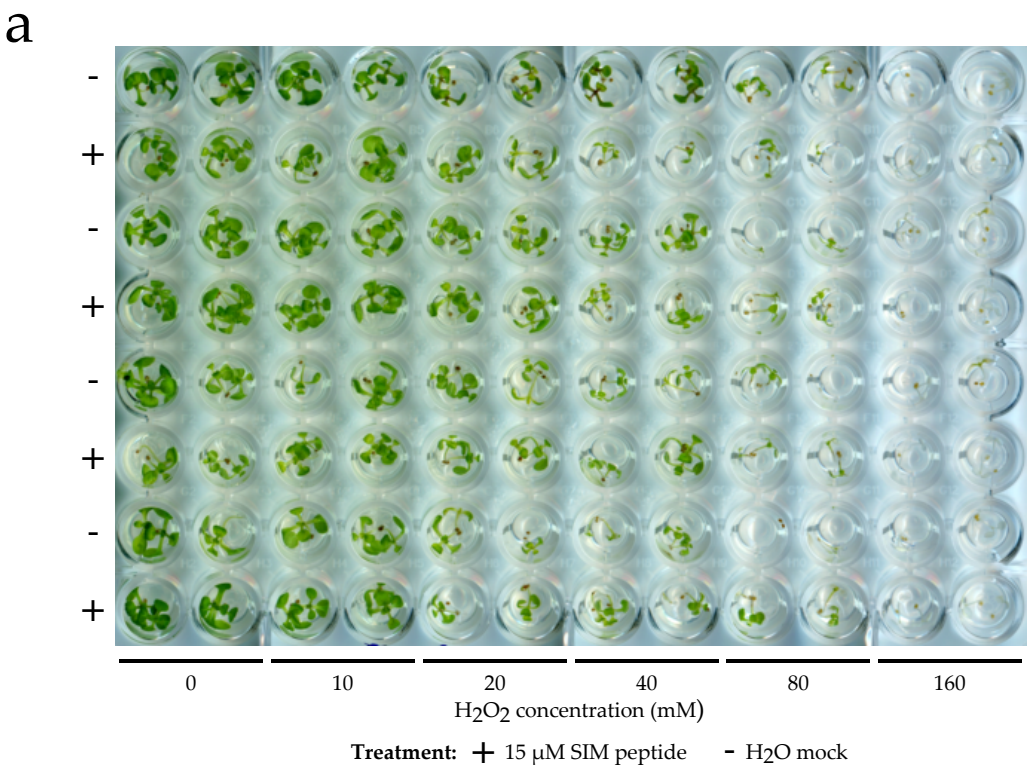
SUMOylation plays an important regulatory role in a myriad of cellular processes. It was hypothesised that SUMO regulated processes could be interfered with by introducing large amounts of SUMO binding peptides into plant cells. Short peptides containing SIMs were proposed to act as competitive inhibitors of SUMO-SIM interactions, reducing the interactions between SUMOylated and SUMO binding proteins. The GID1 SIM peptides screened in Figure 5.3c had been shown to interact with AtSUM1 and an experiment was set up to test whether applying large amounts of free SIM peptide to developing *Arabidopsis* seedlings would induce a phenotype caused by dysregulation of SUMO controlled processes. SUMO plays a dominant role in stress responses and levels of SUMOylation increase in response to stresses. Various concentrations of hydrogen peroxide (H_2O_2) were used to induce oxidative stress which has been shown to cause accumulation of SUMOylated proteins (Kurepa *et al.*, 2003). It was expected that by interfering with SUMO-SIM interactions, stress responses and repair mechanisms would be compromised and SIM peptide treated plants would display a more severe phenotype to the oxidative stress inducer H_2O_2.

Large amounts of SIM peptides from a number of GID1 homologs were synthesised by Bayer Crop Science. The synthesised peptides are a patented invention of Durham University (2014) [UK patent WO2014083301 (A1)]. Initially the AtGID1a SIM peptide was going to be used in this assay but the purity of this sample was lower than the other peptides that were manufactured. Instead the wheat GID1 SIM peptide was used, which differed by a single amino acid (sequence: VVPLHTWVLISNF) and had been shown to interact with AtSUM1 (Figure 5.3c). *Arabidopsis* seedlings were grown in a 96 well plate with 100 µl of ½ Murashige and Skoog (MS) agar by placing two seeds in each well. 6 days after vernalisation the wells were supplemented with 100 µl 15 µM SIM peptide or H_2O mock treatment then two days later the wells were treated with a series of H_2O_2 concentrations. This regime was designed to allow the plants to first absorb the SIM peptide and allow its effect to become established before challenging the plants with the oxidative stress agent. After 10 days of H_2O_2 exposure, data were collected from this 96 well plate experiment and size of the plants was estimated by measuring the colour density of the blue channel of an image of the 96 well plate. The plants strongly absorbed blue light giving a good contrast against the agar and 96 well plate and this was used as a proxy for plant size.

The results from the assay in Figure 5.7 show the expected decline in plant size with increasing H_2O_2 levels, with the highest concentration killing and bleaching the plants. The peptide treatments however, did not show any effect at any concentration of H_2O_2 suggesting that in this assay, the capacity of the plants to cope with oxidative stress had not been compromised.

While the lack of observed effect could be due to the SIM peptide not having an effect on cellular function, it could also be due to insufficient peptide entering and/or remaining within the plant cells

to elicit an effect. The SIM peptides are relatively large, high molecular weight molecules and uptake through the plant tissues and into the cells may be inhibited by the size of these molecules. Additionally the stability of the peptide both internally and externally is not known and the peptide may be degraded by a number of processes including by endogenous peptidases. If the SIM peptide was degraded, the half life of the molecule may be too short to observe an effect. Cellular uptake and peptide stability would need to be investigated before any definitive conclusions can be made about the effect of introducing SIM binding peptide into a plant cell.



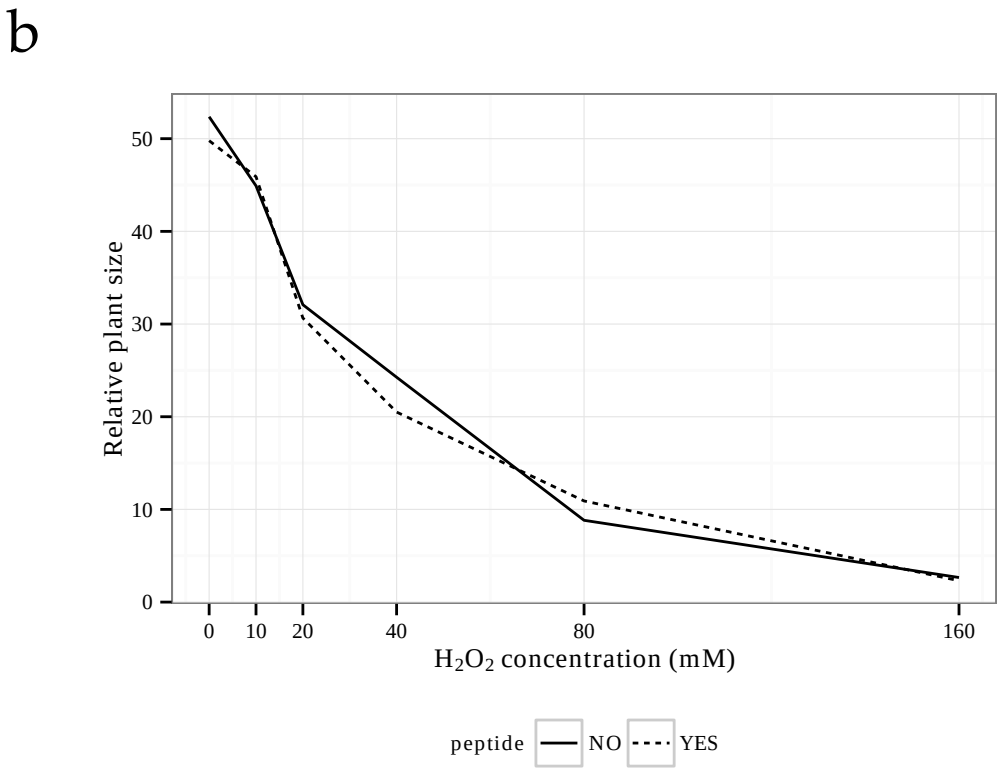


Figure 5.7: Synthetic SIM peptide does not affect plant growth under stress. Plants treated with H_2O_2 show expected decrease in size but treatment with SIM peptide has no effect. **(a)** Two *Arabidopsis* plants were planted on 100 µl of ½ MS agar and allowed to germinate. The wells were treated with a various H_2O_2 concentrations and either 15 µM SIM peptide or a water mock treatment. **(b)** Graph of plant size estimated using the inverse blue value for each well in the 96 well plate image. $n = 8$.

5.3.7 Phenotype of GID1a SIM mutants in *Arabidopsis*

Once the GID1a SIM B peptide with the V8A mutation had been shown to be a weak interactor of AtSUM1 and that the mutation introduced into the full length GID1a protein, gid1a V22A, was still able to bind to RGA in a GA dependant manner, the effect of expressing these two proteins was investigated. The coding sequences (CDS) for *GID1a* and *gid1a V22A* were subcloned from pENTR D/TOPO into the plant expression vector pEarlyGate 201 to generate N-terminal HA fusion genes under the strong 35S promoter of the cauliflower mosaic virus. These vectors were transformed into the Col-0 *Arabidopsis* ecotype and into the SUMO protease mutant *ots1 ots2* using the floral dip method (*ots1 ots2* knock-downs were confirmed by PCR, Figure A.1 in the Appendices). The *ots1 ots2* line was included as these plants hyper-accumulate SUMOylated proteins and it was hypothesised that overexpression of the GID1a proteins in these lines may produce a stronger phenotype. During screening the *ots1 ots2* lines, overexpression of *35S:HA:gid1a V22A* resulted in a severe dwarf phenotype and in these plants the development of the majority of siliques terminated before seeds were produced. Additionally no homozygous lines for *ots1 ots2 35S:HA:GID1a* were isolated in screening. Furthermore, in both the *35S:HA:gid1a V22A* homozygous T_4 lines in the *ots1 ots2* background, there were no detectable levels of the transgenic proteins (Figure 5.8). Expression levels for *ots1 ots2 35S:HA:GID1a* were not tested as there were no homozygous lines for this genotype. Work on hemizygous T_2 *ots1 ots2* transgenic lines published in Conti *et al.* (2008) showed protein expression of the *GID1a* and *gid1a V22A* transgenes. It is likely that the transgenes have been silenced in the T_4 homozygous *ots1 ots2 35S:HA:gid1a V22A* lines which would explain the lack of expression. The transgenes in the wild-type Col-0 lines on the other hand did not have the issues observed with *ots1 ots2* lines and therefore only overexpressors in the wild-type background were investigated further.

Phenotyping the germination rate of *GID1a* and *gid1a V22A* overexpressing lines was performed on both hemizygous and homozygous lines. The germination assay using the hemizygous lines was performed as a preliminary screen which was followed up using seeds from homozygous lines. The results of both experiments are presented as the effects between the hemizygous and homozygous lines are different, possibly due to higher gene dose effects in the homozygous lines. Equal transgene expression in the lines used was confirmed by western blot of the plant extracts (Figure 5.8). Lines 1 and 2 for *35S:HA:GID1a* and lines 2 and 3 for *35S:HA:GID1a V22A* showed similar expression and were used for the germination assay.

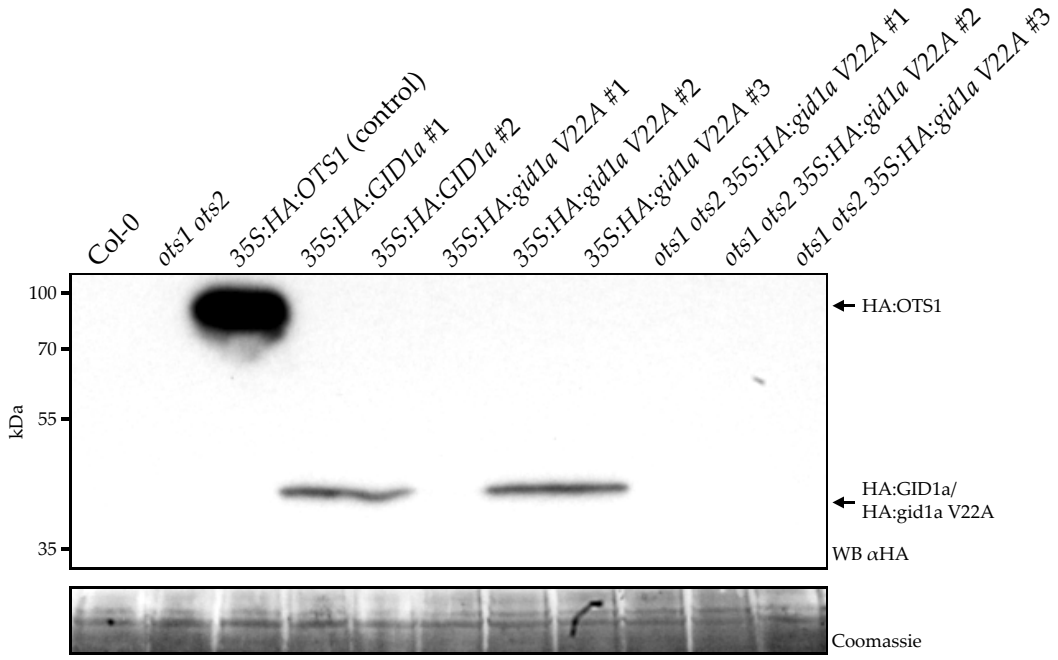


Figure 5.8: Protein expression analysis of transgenic GID1a and gid1a V22A overexpressing plant lines. Two lines for each construct, *35S:HA:GID1a* and *35S:HA:gid1a V22A*, were identified with similar expression levels in the wild-type Col-0 background. No transgene expression was detected in the *ots1 ots2* background lines. Coomassie staining of the blot shows equal protein loading in each lane. WB: western blot.

GA signalling positively regulates seed germination and the effect of the overexpressing transgenes on the germination rate was investigated. Germination rate was used as a proxy for GA signalling as it is regulated by this hormone. It was hypothesised that by overexpressing GID1a the rate of DELLA protein degradation would be increased leading to lower amounts of these transcriptional repressors. GA signalling would be enhanced with lower DELLA levels which would lead to a higher germination rate. A similar effect was hypothesised for the gid1a V22A lines but the effect was expected to be larger thus there would be a higher germination rate. The gid1a V22A mutant protein is expected to have a lower affinity to SUMOylated DELLA proteins which were predicted to inhibit GID1 function. As the gid1a V22A mutant was predicted to be less inhibited by SUMOylated DELLA proteins it was predicted to target more DELLA proteins for degradation compared to overexpression of GID1a alone. The gid1a V22A overexpressing lines were then expected to have enhanced GA signalling compared to overexpressing GID1a lines. The GA biosynthesis inhibitor paclobutrazol (PAC) was included at two concentrations to lower the levels of endogenous GA in the germinating seeds. PAC treatment of germinating seeds leads to lower germination rates (Lee *et al.*, 2002) and this inhibitor was used to test whether the overexpression of GID1a and gid1a V22A could reverse the lower germination phenotype observed with PAC treatment.

The germination rates of second generation transgenic (T_2) seeds were tested on PAC concentrations of 0, 0.1 and 0.5 μM. The seeds had been collected from T_1 parents with a single transgene insertion

event and seeds contained a mix of segregating no-transgene:hemizygous:homozygous seeds with an expected ratio 1:2:1 so the observed phenotype resulted from a mix of those genotypes. The results of the assay using the segregating seeds are shown in Figure 5.9 which shows a slightly lower germination rate for the GID1a overexpressing lines. However, the trend for line GID1a #1 is reverse to what was expected with increasing PAC levels and this result was considered to be anomalous. Comparing the other GID1a overexpressing line #2 to the wild-type shows no significant difference. Overexpression of the gid1a V22A protein on the other hand led to significantly higher germination rates compared to both the GID1a overexpressing lines and to the wild-type (Chi squared test with post-hoc analysis; $p < 0.05$).

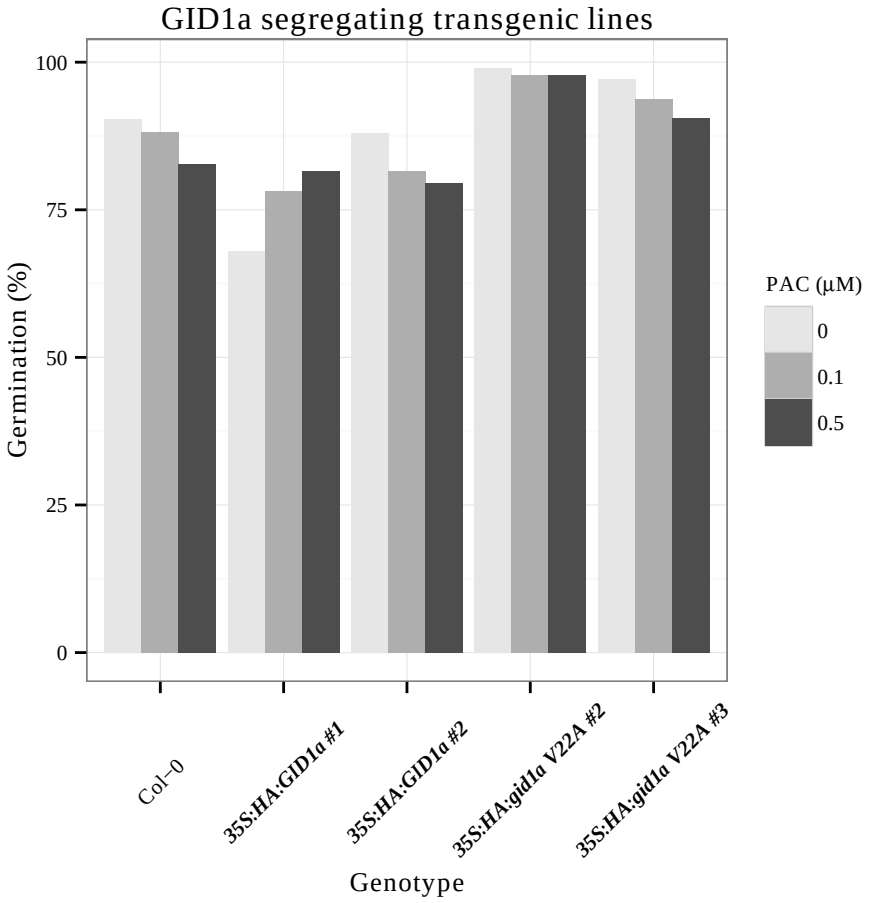


Figure 5.9: Germination rates for over-expressing GID1a and gid1a V22A segregating lines. One *35S:HA:GID1a* line shows lower germination rates to the control while the other line is similar. The *35S:HA:gid1a V22A* lines show significantly higher germination rates compared to the wild-type control and the *35S:HA:GID1a* lines (Chi squared test with post hoc analysis; $p < 0.05$). The results for *35S:HA:GID1a* line #1 are anomalous as they display the reverse trend expected for increasing PAC concentrations, the reason for this anomaly in not known. $n \approx 200$ - number of seeds used for each group.

This assay was repeated again once the homozygous lines had been isolated. The effect of the transgene was expected to be stronger due to a higher overall gene dose effect and that all seeds were transgenic. The results for this assay are shown in Figure 5.10 and both the GID1a and the gid1a V22A

overexpressing lines have close to 100% germination under all concentrations of PAC, significantly higher than the wild-type control ($p < 0.05$). Furthermore, the two different transgenic constructs are not distinguishable from each other, nor were the germination rates between the different PAC treatment levels for the transgenic lines. This suggests that over-expression of both GID1a and gid1a V22A is able to reverse the decrease in germination rate caused by PAC. Interestingly, to further support these results, one of the lines which showed no protein expression (*35S:HA:gid1a V22A* #1; Figure 5.8) was tested and showed similar germination rates to the Col-0 wild-type line (data not shown) demonstrating that observed effects can be attributed to the presence of the overexpressed transgenes.

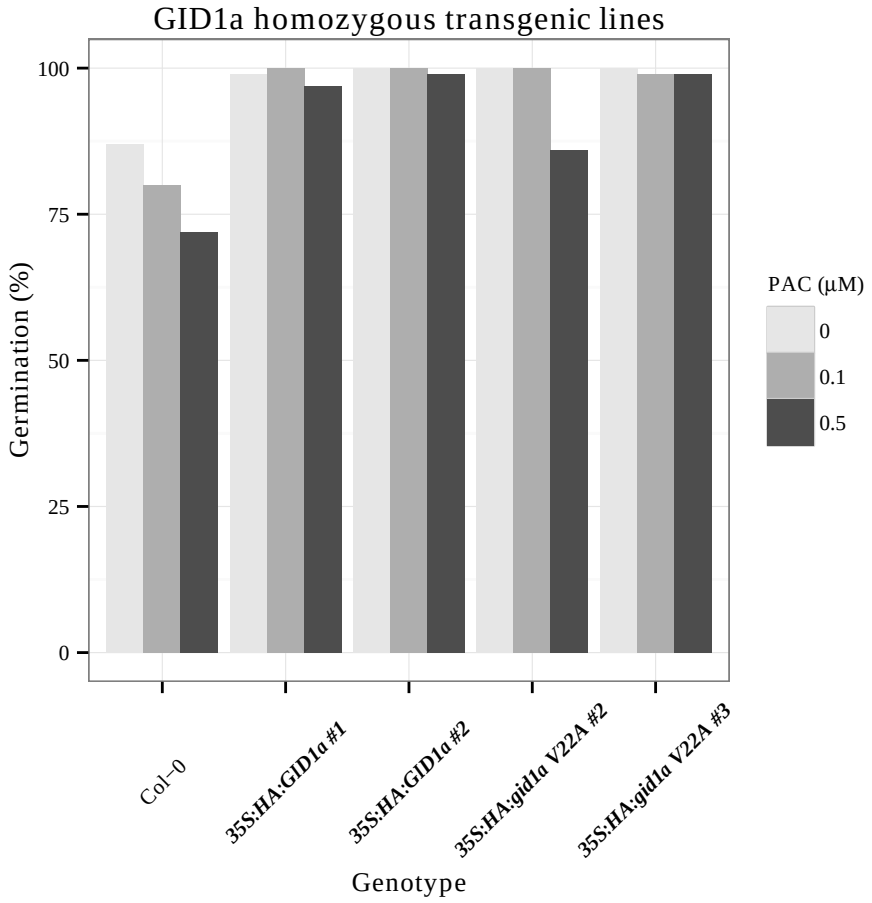


Figure 5.10: Germination rates for over-expressing GID1a and gid1a V22A homozygous lines. All transgenic lines show germination rates close to 100% for all PAC treatments and the germination rates are significantly higher than the wild-type control Col-0 (Chi squared test with post-hoc analysis; $p < 0.05$). $n = 100$ - number of seeds used for each group.

Taking the homozygous lines alone, no difference between the two transgenes can be observed but the data from the segregating lines suggests that the effect of the GID1a overexpressing lines is less than that of gid1a V22A lines. Taking both experiments into consideration it could be concluded that overexpression of gid1a V22A has a strong effect on increasing germination rates while overexpression of GID1a has an intermediate effect between wild-type seeds and gid1a V22A overexpressing seeds.

The germination experiment using the homozygous lines did not show consistent differences between the different PAC treatments and the use of higher concentrations would be required to show differences between the treatment levels and confirm whether the effect of overexpression of GID1a is intermediate between wild type and gid1a V22A.

5.4 Discussion

The purpose of this work was to investigate whether the GA receptor GID1 contained a SIM and could bind to SUMO to test the model of GA independent signalling proposed by Conti *et al.* (2014). This model proposes that the DELLA proteins are SUMOylated under stress conditions and these SUMOylated forms of DELLA are able to bind to and inhibit the action of GID1 and that the mechanism is GA independent. In Chapter 4 it was demonstrated that the *Arabidopsis* DELLA protein RGA is SUMOylated at K65 and this is the only site of SUMOylation in *in vitro* experiments. Research conducted by other colleagues demonstrated that SUMOylated RGA purified from plants could interact with GID1a in a GA independent manner using a co-IP assay; the same experiment demonstrated that the unmodified form of this protein required GA for the interaction with GID1a to occur (Conti *et al.*, 2008) supporting the SUMOylation model of GA independent signalling.

The work described in this chapter complemented the results from the SUMOylated RGA interaction experiment by demonstrating that GID1 can interact with free SUMO, as the model predicted that SUMOylated DELLAs would interact with GID1 via binding of the SUMO group attached to the DELLA proteins to a SIM or group of SIMs in GID1. Bioinformatic analysis suggested that there was at least one SIM in the lid domain of GID1 and the predicted regions were highly conserved in both *Arabidopsis* paralogs and in orthologs in cereal plant species. Interaction with these predicted SIM regions was confirmed using short peptides of these sequences. A single region named SIM B covering residues V15 to F27 in the GID1 proteins tested interacted with AtSUM1 and peptides from all species were shown to interact with AtSUM1. These results suggested that the SUMO binding capacity of this region is conserved across a wide evolutionary range of plant species. Later analysis using the HyperSUMO sequence feature predictor strongly suggested that a hydrophobic tetrad required for SUMO interaction consisted of the first four residues of SIM B (V15-L18) and that the SIM was of type A with the important charged or polar amino acids lying upstream of the hydrophobic core. SIM B lies within the lid domain of GID1, within an unstructured loop region and part of α-helix B. Generally SIMs cannot be part of secondary structures such as α-helices, however, the structure of GID1a was resolved using the hormone and DELLA bound form of the protein and it is likely that the protein takes on a different structure in the unbound state. Simulation of the dynamics of GID1a in the unbound state, GID1a-GA_4 and GID1a-GA_4-DELLA has shown that the secondary structure of α-helix B in GID1a is stabilised by

both the binding of GA and the DELLA protein and suggests that in the unbound state the lid domain of GID1a has much greater conformational flexibility (Hao *et al.*, 2013). Given the conformational flexibility of the GID1a α-helix B in the unbound state, it is plausible that SUMO binding can occur. These structural data suggest that the hormone bound state of GID1 may not be suitable for SUMO binding to SIM B but rather binding occurs in the conformation of the unbound state.

Attempts to generate mutant SIM versions which lacked SIM binding were only partially successful, identifying the weakly interacting SIM B V8A peptide. The inability to generate complete loss of interaction mutants was probably due to targeting the incorrect amino acids for mutagenesis, as the peptide sequence at positions 7W-10I (respective to the position in the peptide, not GID1a) were originally thought to form the hydrophobic core of the SIM. Later, using results from the work described in Chapter 3 this hypothesis was updated to residues 1V-L4 instead. Targeting these residues for mutagenesis instead may have resulted in complete loss of SUMO interaction. The weakly interacting V8A mutant was investigated further and introduced into the full length sequence of a *GID1a* clone, generating *gid1a V22A*. A Y2H screen confirmed that the gid1a V22A mutant was functional and could bind to RGA in a GA dependent manner, though the interaction appeared to be weaker. An alternative mutant was also generated, gid1a V22S, and this mutant too was able to bind to RGA in a GA dependent manner but the strength of the interaction was lower than that of gid1a V22A. The ability of the gid1a V22S mutant to interact with SUMO was not confirmed.

Although the mutagenised SIM B V8A peptide had been shown to be a weaker interactor of AtSUM1, the interaction of the corresponding mutation in GID1a was not tested and remains an important outstanding experiment that needs to be performed in future work. This experiment is required to confirm whether the affinity of full length gid1a V22A protein for AtSUM1 is weaker. Originally it was planned to investigate and measure the binding kinetics of native GID1a and the two SIM mutant forms with both AtSUM1 and RGA using SPR but due to technical issues the assay failed. These data would have confirmed whether the GID1a SIM mutants did indeed have a lower binding affinity for AtSUM1. The SPR binding data was also going to be used to test whether the model of SUMOylated DELLA proteins inhibiting GID1 was plausible from a kinetics point of view by modelling the proposed pathway. Immobilisation of GID1a and the two mutagenised versions to an SPR sensor chip proved to be very inefficient despite screening for optimal pH for binding. For the control GST protein on the other hand, there were no issues with immobilisation.

Though only a small amount of protein from the GID1a samples was immobilised, the interaction assay was performed with AtSUM1 and RGA. The measurements obtained however, appeared to be due to background binding rather than actual GID1a protein interaction as no differential responses were seen between the RGA runs with and without GA. A stronger signal was expected from the GID1a-GA-RGA run than from the GID1a-RGA run as RGA binding has been demonstrated to be GA dependant.

This expected difference in binding was not seen, indicating that the expected molecular interactions were not occurring. One possible explanation could be that the wrong protein species were captured from the heterogeneous GID1a samples which contained a number of break down products and a protein fragment without a functional GID1a domain could have been preferentially captured. Another group investigating the interaction of RGA and AtSUM1 with GID1a were able to show interaction using SPR (Woodcock, 2014). They used protein samples donated by myself for the interaction assay that were expressed and purified in the same manner as the samples described in this chapter. Woodcock (2014) used a lower pH of 4.5 for the immobilisation of GID1a suggesting that the higher pH levels used in the SPR assay described here were not appropriate for capturing GID1a. However, Woodcock (2014) used buffers lacking salts to achieve binding and could not show GA dependance for the GID1a-RGA interaction. It was suggested that the recombinant RGA protein behaved differently to the natural protein and showed uncharacteristic binding to GID1a in the absence of GA. This assumption was made because RGA bound to GID1a displayed a very slow dissociation from GID1a. Overall these results suggest that there remains a major technical issue in the interaction assay that prevents the measurement of the binding kinetics of both AtSUM1 to GID1a and RGA to GID1a.

The issue of degradation of the GST:GID1a fusion, which was used in this assay, was noted by Murase *et al.* (2008) during their preparation of GID1a for crystallisation and they also found that the protein was less stable in the absence of GA. Murase *et al.* (2008) proposed that this was due to more solvent exposed protein chains in the GA free sample of GID1a they used which allowed more favourable conditions for protease digestion. The GID1a samples described in this chapter were expressed and purified in the absence of GA which was likely to have enhanced the rate of degradation of this protein. To alleviate the problems with sample heterogeneity, the removal of the GST tag by protease cleavage at a specific recognition site followed by size exclusion chromatography would allow isolation of a homogeneous sample. Immobilisation of GID1a from such a sample with a single protein species would eliminate the issue of immobilising a degradation fragment.

The observed binding signal in the SPR experiments could be explained by the signal from non-specific interactions which had not been completely subtracted from the final SPR sensogram. The signal from the GST channel was subtracted from the signal from the GID1a, gid1a V22A and gid1a V22S channels but had to be scaled since the protein amounts in each channel were not a 1:1 molar ratio. Rather the GST control was significantly higher. It is possible that due to the large difference in molar ratios and errors in the scaling calculations, the signal from non-specific binding was not completely subtracted.

Although reduced AtSUMO interaction for gid1a V22A full length protein had not been confirmed, overexpression experiments in *Arabidopsis* were carried out which supported reduced SUMO binding. With all other variables being equal, the amount of GID1 protein in the cell determines the stability of the

DELLA proteins. Overexpression of GID1 would be expected to reduce the amount of DELLA proteins and enhance GA signalling and this has been demonstrated by a number of groups (Ueguchi-Tanaka *et al.*, 2005; Conti *et al.*, 2014). It has been shown that a small pool of DELLA proteins are always SUMOylated, which increases under stress conditions (Conti *et al.*, 2014). The model of SUMOylated DELLA inhibition of GID1a then predicts that there is always a certain level of GID1 inhibition, which is enhanced under stress conditions. Overexpression of GID1 protein lacking or having reduced SUMO binding would be expected to have a stronger effect on reducing DELLA protein levels as the GID1a inhibitory effect would be reduced. This would lead to enhanced GA signalling compared to overexpressing wild-type GID1a. Based on these hypotheses it was predicted that overexpression of the weak AtSUM1 interacting mutant gid1a V22A would lead to stronger GA signalling than the overexpression of GID1a alone. Germination was used as a proxy for GA signalling as the process is strongly dependant on GA signalling (Ogawa *et al.*, 2003). In concordance with previously published results, the overexpression of GID1a was demonstrated to enhance the germination rate of *Arabidopsis* seeds. The effect was found to be significantly higher for overexpression of gid1a V22A, however, the effect was only observed in segregating lines as the strength of the effect of both homozygous overexpressors led to almost complete germination of the seeds tested. To observe differences in the strength of effect in homozygous overexpressing lines, higher levels of the GA biosynthesis inhibitor PAC could be used to provide more challenging assay conditions.

The results from the overexpression assays support predictions made from the model of SUMOylated DELLA induced GID1 inhibition which predicted enhanced GA signalling in GID1 SIM mutant overexpressors. Although the results from the work described in this chapter support the model and no results have been obtained that disprove the model, there are a number of critical aspects that have not been tested. Most importantly the model predicts that SUMOylated DELLA is responsible for GID1 inhibition, however, it has not been demonstrated experimentally that it is actually the case and alternatively there could be some other SUMOylated target other than the DELLA proteins that is inhibiting GID1. Stress induced SUMOylation results in the SUMOylation of a large number of protein targets and one or more of these could be inhibiting GID1. Determining the identity of the SUMOylated target which inhibits GID1 activity should be the next area of research into the role of SUMOylation in GA signalling.

Chapter 6

Discussion

The purpose of the work presented in the thesis was to expand the understanding of SUMO-SIM interactions in plants and to develop a set of tools to predict SIMs in protein sequences. Two separate areas of research were presented in this thesis, the first was a large investigation into the amino acid composition of SIMs that bind to AtSUM1 andHsSUM1 while the latter part investigated a specific role of a SUMO-SIM interaction in gibberellin receptor GID1.

The results of this work demonstrate distinct differences in the binding properties of plant and animal SUMO proteins. Phosphorylation of the SIMs was shown to have either an activating or a deactivating effect in SIMs depending on the location of the modification and the amino acid composition of the SIM. Large datasets of SIM interactions with both AtSUM1 and HsSUM1 were generated and the data were used to construct SIM predictors using random forest models. A SUMO site predictor was also developed using random forest models and both the SIM and SUMO site predictors were combined to develop a web-based SUMO-related sequence feature predictor with a graphical user interface to make the predictor available to the wider research community. The plant SIM predictor that was developed was used to perform a large-scale screen for conserved SUMO binding proteins (SBPs) in *Arabidopsis*. The predicted SBPs form a group of novel proteins whose function could be regulated by SUMO and these proteins are good candidates for further research into the role of SUMOylation. The predicted SBPs were enriched for biological processes that are known to be regulated by SUMO, such as DNA and RNA maintenance as well as for novel processes such as cell wall metabolism.

Investigation into the role of SUMO in the GA pathway was initiated by the finding that the DELLA repressor protein RGA is a target for SUMOylation. A model was proposed whereby SUMOylated RGA inhibits the action of the GA receptor GID1 which targets the DELLA proteins for degradation in the presence of GA. SUMOylated DELLA inhibition of GID1 was proposed to occur through binding of the SUMO moiety of SUMOylated RGA to GID1 via a SIM, blocking the action of GID1. The results presented in Chapter 5 demonstrate that GID1 can bind to SUMO and that the model of SUMOylated DELLA inhibition of GID1 is plausible.

6.1 Analysis of SIM sequences

Research in animal models has shown that SUMO isoforms have different affinities for SIMs depending on the amino acid sequence. SUMO isoform binding differences were observed between human isoforms of SUMO (Tatham *et al.*, 2005) so inter-species differences were expected to exist. To what extent plant and human SUMO isoforms differed in their binding properties was unknown prior to this work. Though plant SBPs had been identified, prior to the publication of the research described in this thesis by Conti *et al.* (2014), no plant SIMs had been described so no data was available for characterising plant SUMO binding preferences.

In order to investigate the binding properties of plant SUMO, a randomly generated peptide library was designed to identify AtSUM1 binding peptides. Namanja *et al.* (2012) used a large-scale peptide library to investigate the binding differences between HsSUM1 and HsSUM3, though their approach is different to that taken in this work. For two known human SIMs, Namanja *et al.* (2012) generated all possible single amino acid substitutions of 13-residue sequences and used these sequences to generate peptide arrays. Although this peptide library screened hundreds of SIMs, they were all very similar, with any two peptides differing by at most two amino acids. For the purpose of their work, this approach was appropriate as they were interested in identifying important amino acid residues within those specific SIMs. However, using the same strategy to generate *de novo* data to characterise SUMO binding properties would result in a very constrained dataset that would not sample very much peptide sequence variety. To overcome this issue, random sequences from a defined distribution were generated. The distribution that the SIMs were drawn from was constrained to a limited set of amino acids at certain positions to ensure that the peptide resembled the general motif of known SIMs. Three types of SIM motif are known (Vogt & Hofmann, 2012) and peptides resembling each type were generated. The advantage of this approach was that the peptides generated had a greater variety and sampled a more diverse range of peptide sequences. Although there was no replication of the peptides on the arrays, there was position specific replication of amino acids in the peptide library, which allowed trends in amino acid binding preferences to be observed.

The constraints imposed upon the amino acids at the different peptide positions were quite stringent and limited the variety of the peptides in the set. Importantly only a limited number of amino acids were screened at the most constrained hydrophobic core positions of the peptides and it is possible that some features of SIMs could have been missed using this dataset. The stringency was set high to ensure that enough interacting peptides were identified. A peptide library with higher diversity would sample more peptide diversity but have a lower proportion of interacting peptides and due to limited library size, too few SUMO interactors would have been identified to be useful. A balance between peptide variety and positive interactions had to be met.

The approach of using a randomly generated SIM-like peptide library produced an acceptable ratio of interacting to non-interacting peptides, allowing trends in the peptide sequences to be observed. AtSUM1 binding preferences were compared with those of HsSUM1, and the binding properties of the two SUMO isoforms were found to be very different with only a minority of the library peptides being able to bind to both SUMO isoforms. The major differences in binding were found to be due to the amino acids immediately next to the hydrophobic cores of the SIMs, with the influence of the position on binding having less of an effect the further from the core it was. Furthermore, phosphorylation of polar amino acids within certain SIMs was shown to act as interaction switches either activating or deactivating the interaction. In a biological context, this is an important observation showing that effects

of SUMOylation can be regulated by protein kinase cascades.

These results demonstrate that there are large differences in SUMO binding behaviour in distantly related species. The peptide sequence requirements for SUMO isoforms from distantly related species needs to therefore be determined independently as the binding preferences vary significantly. Furthermore, these results show that any predictors of SIMs will need to be trained using species and SUMO paralog specific data. This is in contrast to the prediction of SUMO sites that do not show species or SUMO isoform differences.

6.2 SUMO-related sequence predictor

The large SIM datasets were used to build specific SIM predictors for both AtSUM1 and HsSUM1 isoforms. Random forest methods had been used previously to predict SUMO sites within proteins by Teng *et al.* (2012). Areas for improvement in these authors' methodology were identified to improve the performance of the predictors. A new SUMO site predictor was built using training data from Xue *et al.* (2006), using the improved random forest method. The performance of the SUMO site predictor was significantly improved by this new method. This increase in performance was achieved by converting amino acid factor variables into numeric PCA dimensions, by performing variable selection and by performing parameter optimisation.

The resulting SUMO site predictors outperformed previously published predictors and offered both better sensitivity and specificity. The SIM predictors had lower performance because of limited training data. The performance of the SIM predictors could be improved by using training datasets with more observations. Future work should focus on building larger SIM datasets and should include additional SUMO paralogs from *Arabidopsis* including AtSUM2 and AtSUM3. AtSUM1 and AtSUM2 share a high sequence similarity (Castaño-Miquel *et al.*, 2011) and it would interesting to see to what degree, if any, the SIM binding preferences differ between these two proteins. Future work could also use quantitative interaction methods, such as using fluorescently labelled SUMO proteins. This would allow more complex models of SUMO interaction to be incorporated into a SIM predictor. Chapter 3 discusses some other minor technical issues with the peptide synthesis quality and methods for addressing these issues.

The AtSUM1 SIM predictor, along with structural and evolutionary information, was used to identify putative SBPs containing at least two predicted SIMs in *Arabidopsis* using a genome wide screen of sequences. The screen focused on evolutionarily conserved proteins having predicted orthologous proteins in other plant species. The SBPs identified in this screen agree well with the role of SUMOylation in DNA and RNA maintenance and repair (Elrouby *et al.*, 2013; Xu *et al.*, 2013; Mazur & van den Burg, 2012). While the false positive rate for the predicted SBPs is unknown, this set of pro-

teins is expected to contain an enrichment of true SBPs, as the set contains the expected enrichment for DNA/RNA related processes. The identified proteins complement work by Miller *et al.* (2012) who have used genome wide screens to identify protein targets for SUMOylation, and some of these SUMOylated proteins may be the interacting partners for the predicted SBPs. Co-localisation and co-expression studies of the identified SUMOylated proteins and predicted SBPs could be used to identify putative interacting SBP-SUMOylated protein pairs. There is also likely to be some overlap in the predicted SBPs and other large scale screens for SUMOylated proteins. SUMOylated proteins often contain SIMs which promote SUMOylation of these proteins (Jentsch & Psakhye, 2013) and these proteins could also be identified in the SBP screen.

6.3 The role of SUMOylated DELLAs

DELLA proteins in the GA pathway are responsible for negatively regulating GA responses in the absence of the hormone GA. In the presence of GA, the GA receptor GID1 binds to the DELLA proteins and targets them for degradation leading to the activation of GA signalling. The demonstration that the DELLA proteins are SUMOylated (Conti *et al.*, 2014) was the first evidence for the role of SUMO in the GA pathway.

The model of SUMOylated DELLA inhibition of the receptor GID1a through the binding of SUMOylated DELLA to GID1 via the SUMO moiety was proposed as the function for this modification. In this model, increased levels of SUMOylated DELLA proteins lead to more binding of SUMOylated DELLAs to GID1. These SUMOylated DELLA proteins are not targeted for degradation by GID1, and the binding of SUMOylated DELLA inhibits the activity of GID1 against non-modified DELLAs. This model then predicts that when SUMOylated levels of DELLA proteins increase, such as in response to stress, inhibition of GID1 occurs allowing higher levels of DELLA proteins to accumulate, which in turn represses GA signalling and restrains plant growth. It is predicted that increasing the levels of GID1 through overexpression would reduce DELLA levels, and enhance GA signalling, leading to increased growth. Furthermore, overexpression of a GID1 mutant protein lacking SUMO binding would be expected to have a stronger effect on reducing DELLA protein levels as it would not be subject to inhibition by the pool of SUMOylated DELLAs.

Results from Chapter 5 demonstrated that GID1a protein is able to bind to SUMO and a putative SIM was identified in the lid region of the protein. Co-IP assays performed by another collaborating researcher demonstrated that SUMOylated RGA could bind to GID1a in a GA independent manner (Conti *et al.*, 2014). These experiments demonstrated that SUMOylated DELLA binding to GID1a could occur *in vitro.* Overexpression lines were used to test the two hypotheses regarding overexpression of GID1 and a GID1 SIM mutant. Germination assays confirmed that overexpression of a SIM mutant of GID1,

gid1 V22A, was able to increase seed germination rate, which was used as a proxy for GA signalling. The results for overexpression of GID1a alone were inconclusive, as an assay using hemizygous lines did not show an increase in germination rate, while in an assay using homozygous lines overexpressing GID1a and gid1a V22A resulted in almost complete germination of the seeds tested, even when exposed to the GA biosynthesis inhibitor PAC. Nevertheless these plant overexpression experiments support the proposed hypotheses regarding GA signalling.

Although a number of experiments have provided evidence for an inhibitory effect of SUMOylated RGA on GID1a activity, no direct evidence of the mechanism has been demonstrated *in planta*. The interaction between SUMOylated DELLA and GID1a needs to be tested in plant cells to further test this model. Förster resonance energy transfer (FRET) is commonly used to demonstrate the interaction of two proteins *in vivo* and recent advances in methodology (Wallrabe *et al.*, 2012) and results analysis (Hoppe *et al.*, 2013) have allowed interaction of more than two proteins to be investigated. Each protein is fused to a different fluorescent molecule and each pair of possible interactions can be tested by exciting different fluorophores and observing the emission spectra. Using three different fluorophores to tag a DELLA protein, SUMO1 and GID1a would allow the association of the three molecules to be tested. Demonstrating interaction of these three proteins *in vivo* would provide important direct evidence for the interaction of SUMOylated DELLA proteins with GID1. The SUMOylation of DELLA proteins may be a critical mechanism regulating growth restraint during periods of stress. Understanding of plant hormonal regulation in *Arabidopsis* could inform research into developing new crop plant biotechnologies.

6.4 Concluding remarks

The research presented in this thesis has provided the first detailed investigation of SIMs in plant models. The data from the SIM work was used to build a predictor for SIMs that bind to both AtSUM1 and HsSUM1 and will be of use to researchers investigating SUMOylation in both plant and animal models. A genome-wide screen for SIM containing proteins in *Arabidopsis* has provided a novel set of predicted SBPs, which may help elucidate the role of SUMOylation alongside previous work identifying SUMOylated proteins by other groups. Finally, the role of SUMOylation of the DELLA proteins was investigated and a functional SIM in the GA receptor was identified indicating that the GA pathway is regulated by SUMO.

Bibliography

Achard, P., Cheng, H., De Grauwe, L., Decat, J., Schoutteten, H., Moritz, T., Van Der Straeten, D., Peng, J. & Harberd, N.P. (2006) Integration of plant responses to environmentally activated phytohormonal signals. *Science*, **311**(5757): 91–94.

Alonso-Ramírez, A., Rodríguez, D., Reyes, D., Jiménez, J.A., Nicolás, G., López-Climent, M., Gómez-Cadenas, A. & Nicolás, C. (2009) Evidence for a role of gibberellins in salicylic acid-modulated early plant responses to abiotic stress in *Arabidopsis* seeds. *Plant Physiology*, **150**(3): 1335–1344.

Arana, M.V., Marín-de la Rosa, N., Maloof, J.N., Blázquez, M.A. & Alabadí, D. (2011) Circadian oscillation of gibberellin signaling in *Arabidopsis*. *Proceedings of the National Academy of Sciences of the United States of America*, **108**(22): 9292–9297.

Ariizumi, T., Murase, K., Sun, T.P. & Steber, C.M. (2008) Proteolysis-independent downregulation of DELLA repression in *Arabidopsis* by the gibberellin receptor GIBBERELLIN INSENSITIVE DWARF1. *The Plant Cell*, **20**(9): 2447–2459.

Armstrong, A.A., Mohideen, F. & Lima, C.D. (2012) Recognition of SUMO-modified PCNA requires tandem receptor motifs in Srs2. *Nature*, **483**(7387): 59–63.

Baxter, L., Jironkin, A., Hickman, R., Moore, J., Barrington, C., Krusche, P., Dyer, N.P., Buchanan-Wollaston, V., Tiskin, A., Beynon, J., Denby, K. & Ott, S. (2012) Conserved noncoding sequences highlight shared components of regulatory networks in dicotyledonous plants. *The Plant Cell*, **24**(10): 3949–3965.

Berry, C., Hannenhalli, S., Leipzig, J. & Bushman, F.D. (2006) Selection of target sites for mobile DNA integration in the human genome. *PLoS Computational Biology*, **2**(11): e157.

Böhmdorfer, G., Schleiffer, A., Brunmeir, R., Ferscha, S., Nizhynska, V., Kozák, J., Angelis, K.J., Kreil, D.P. & Schweizer, D. (2011) GMI1, a structural-maintenance-of-chromosomes-hinge domain-containing protein, is involved in somatic homologous recombination in Arabidopsis. *The Plant Journal*, **67**(3): 420–433.

Bolger, G.B., Baillie, G.S., Li, X., Lynch, M.J., Herzyk, P., Mohamed, A., Mitchell, L.H., McCahill, A., Hundsrucker, C., Klussmann, E., Adams, D.R. & Houslay, M.D. (2006) Scanning peptide array analyses identify overlapping binding sites for the signalling scaffold proteins, beta-arrestin and RACK1, in cAMP-specific phosphodiesterase PDE4D5. *The Biochemical Journal*, **398**(1): 23–36.

Bolle, C. (2004) The role of GRAS proteins in plant signal transduction and development. *Planta*, **218**(5): 683–692.

Boyer-Guittaut, M., Birsoy, K., Potel, C., Elliott, G., Jaffray, E., Desterro, J.M., Hay, R.T. & Oelgeschläger, T. (2005) SUMO-1 modification of human transcription factor (TF) IID complex subunits: inhibition of TFIID promoter-binding activity through SUMO-1 modification of hsTAF5. *The Journal of Biological Chemistry*, **280**(11): 9937–9945.

Breiman, L. (1996) Out-of-bag estimation. Technical report, Statistics Department, University of California Berkeley.

Breiman, L. (2001) Random Forests. *Machine Learning*, **45**(1): 5–32.

Budhiraja, R., Hermkes, R., Muller, S., Schmidt, J., Colby, T., Panigrahi, K., Coupland, G. & Bachmair, A. (2009) Substrates related to chromatin and to RNA-dependent processes are modified by Arabidopsis SUMO isoforms that differ in a conserved residue with influence on desumoylation. *Plant Physiology*, **149**(3): 1529–1540.

Burroughs, A.M., Balaji, S., Iyer, L.M. & Aravind, L. (2007) Small but versatile: the extraordinary functional and structural diversity of the beta-grasp fold. *Biology Direct*, **2**: 18.

Butt, T.R., Edavettal, S.C., Hall, J.P. & Mattern, M.R. (2005) SUMO fusion technology for difficult-to-express proteins. *Protein Expression and Purification*, **43**(1): 1–9.

Cai, Q., Cai, S., Zhu, C., Verma, S.C., Choi, J.Y. & Robertson, E.S. (2013) A unique SUMO-2-interacting motif within LANA is essential for KSHV latency. *PLoS Pathogens*, **9**(11): e1003,750.

Castaño-Miquel, L., Seguí, J. & Lois, L.M. (2011) Distinctive properties of *Arabidopsis* SUMO paralogues support the *in vivo* predominant role of AtSUMO1/2 isoforms. *The Biochemical journal*, **436**(3): 581–590.

Castro, P.H., Tavares, R.M., Bejarano, E.R. & Azevedo, H. (2012) SUMO, a heavyweight player in plant abiotic stress responses. *Cellular and Molecular Life Sciences*, **69**(19): 3269–3283.

Catala, R., Ouyang, J., Abreu, I.A., Hu, Y., Seo, H., Zhang, X. & Chua, N.H. (2007) The *Arabidopsis* E3 SUMO ligase SIZ1 regulates plant growth and drought responses. *The Plant Cell*, **19**(9): 2952–2966.

Cha, J.Y., Ahn, G., Kim, J.Y., Kang, S.B., Kim, M.R., Su'udi, M., Kim, W.Y. & Son, D. (2013) Structural and functional differences of cytosolic 90-kDa heat-shock proteins (Hsp90s) in *Arabidopsis thaliana*. *Plant Physiology and Biochemistry*, **70**: 368–373.

Chang, C.C., Naik, M.T., Huang, Y.S., Jeng, J.C., Liao, P.H., Kuo, H.Y., Ho, C.C., Hsieh, Y.L., Lin, C.H., Huang, N.J., Naik, N.M., Kung, C.C.H., Lin, S.Y., Chen, R.H., Chang, K.S., Huang, T.H. & Shih, H.M. (2011) Structural and functional roles of Daxx SIM phosphorylation in SUMO paralog-selective binding and apoptosis modulation. *Molecular Cell*, **42**(1): 62–74.

Charif, D. & Lobry, J. (2007) SeqinR 1.0-2: a contributed package to the R project for statistical computing devoted to biological sequences retrieval and analysis. In *Structural approaches to sequence evolution: Molecules, networks, populations* (edited by U. Bastolla, M. Porto, H. Roman & M. Vendruscolo), Biological and Medical Physics, Biomedical Engineering. Springer Verlag, New York, pp. 207–232. ISBN : 978-3-540-35305-8.

Chen, C.C., Chen, Y.Y., Tang, I.C., Liang, H.M., Lai, C.C., Chiou, J.M. & Yeh, K.C. (2011) *Arabidopsis* SUMO E3 ligase SIZ1 is involved in excess copper tolerance. *Plant Physiology*, **156**(4): 2225–2234.

Chen, I.P., Mannuss, A., Orel, N., Heitzeberg, F. & Puchta, H. (2008) A homolog of ScRAD5 is involved in DNA repair and homologous recombination in Arabidopsis. *Plant Physiology*, **146**(4): 1786–1796.

Chosed, R., Mukherjee, S., Lois, L.M. & Orth, K. (2006) Evolution of a signalling system that incorporates both redundancy and diversity: *Arabidopsis* SUMOylation. *The Biochemical Journal*, **398**(3): 521–529.

Clough, S.J. & Bent, A.F. (1998) Floral dip: a simplified method for *Agrobacterium*-mediated transformation of *Arabidopsis thaliana*. *The Plant Journal*, **16**(6): 735–743.

Conti, L., Nelis, S., Zhang, C., Woodcock, A., Swarup, R., Galbiati, M., Tonelli, C., Napier, R., Hedden, P., Bennett, M. & Sadanandom, A. (2014) Small Ubiquitin-like Modifier protein SUMO enables plants to control growth independently of the phytohormone gibberellin. *Developmental Cell*, **28**(1): 102–110.

Conti, L., Price, G., O'Donnell, E., Schwessinger, B., Dominy, P. & Sadanandom, A. (2008) Small ubiquitin-like modifier proteases OVERLY TOLERANT TO SALT1 and -2 regulate salt stress responses in *Arabidopsis*. *The Plant Cell*, **20**(10): 2894–2908.

Da Silva-Ferrada, E., Xolalpa, W., Lang, V., Aillet, F., Martin-Ruiz, I., de la Cruz-Herrera, C.F., Lopitz-Otsoa, F., Carracedo, A., Goldenberg, S.J., Rivas, C., England, P. & Rodríguez, M.S. (2013) Analysis of SUMOylated proteins using SUMO-traps. *Scientific Reports*, **3**: 1690.

Dill, A., Thomas, S.G., Hu, J., Steber, C.M. & Sun, T.P. (2004) The *Arabidopsis* F-box protein SLEEPY1 targets gibberellin signaling repressors for gibberellin-induced degradation. *The Plant cell*, **16**(6): 1392–1405.

Duda, D.M., van Waardenburg, R.C.A.M., Borg, L.A., McGarity, S., Nourse, A., Waddell, M.B., Bjornsti, M.A. & Schulman, B.A. (2007) Structure of a SUMO-binding-motif mimic bound to Smt3p-Ubc9p: conservation of a non-covalent ubiquitin-like protein-E2 complex as a platform for selective interactions within a SUMO pathway. *Journal of Molecular Biology*, **369**(3): 619–630.

Durham University (2014) *Transgenic plants with altered sumoylation*. UK Patent: WO2014083301 (A1).

Earley, K.W., Haag, J.R., Pontes, O., Opper, K., Juehne, T., Song, K. & Pikaard, C.S. (2006) Gateway-compatible vectors for plant functional genomics and proteomics. *The Plant Journal*, **45**(4): 616–629.

Edgar, R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, **32**(5): 1792–1797.

Elrouby, N., Bonequi, M.V., Porri, A. & Coupland, G. (2013) Identification of *Arabidopsis* SUMO-interacting proteins that regulate chromatin activity and developmental transitions. *Proceedings of the National Academy of Sciences of the United States of America*, **110**(49): 19,956–19,961.

Elrouby, N. & Coupland, G. (2010) Proteome-wide screens for small ubiquitin-like modifier (SUMO) substrates identify *Arabidopsis* proteins implicated in diverse biological processes. *Proceedings of the National Academy of Sciences of the United States of America*, **107**(40): 17,415–17,420.

Escobar-Cabrera, E., Okon, M., Lau, D.K.W., Dart, C.F., Bonvin, A.M.J.J. & McIntosh, L.P. (2011) Characterizing the N- and C-terminal Small ubiquitin-like modifier (SUMO)-interacting motifs of the scaffold protein DAXX. *The Journal of Biological Chemistry*, **286**(22): 19,816–19,829.

Fawcett, T. (2006) An introduction to ROC analysis. *Pattern Recognition Letters*, **27**(1): 861–874.

Finkelstein, R.R. & Lynch, T.J. (2000) The *Arabidopsis* abscisic acid response gene emphABI5 encodes a basic leucine zipper transcription factor. *The Plant Cell*, **12**(4): 599–609.

Frank, R. (2002) The SPOT-synthesis technique. Synthetic peptide arrays on membrane supports–principles and applications. *Journal of Immunological Methods*, **267**(1): 13–26.

Fu, X., Richards, D.E., Ait-Ali, T., Hynes, L.W., Ougham, H., Peng, J. & Harberd, N.P. (2002) Gibberellin-mediated proteasome-dependent degradation of the barley DELLA protein SLN1 repressor. *The Plant Cell*, **14**(12): 3191–3200.

Galanty, Y., Belotserkovskaya, R., Coates, J., Polo, S., Miller, K.M. & Jackson, S.P. (2009) Mammalian SUMO E3-ligases PIAS1 and PIAS4 promote responses to DNA double-strand breaks. *Nature*, **462**(7275): 935–939.

Gallego-Bartolomé, J., Minguet, E.G., Grau-Enguix, F., Abbas, M., Locascio, A., Thomas, S.G., Alabadí, D. & Blázquez, M.A. (2012) Molecular mechanism for the interaction between gibberellin and brassinosteroid signaling pathways in *Arabidopsis*. *Proceedings of the National Academy of Sciences of the United States of America*, **109**(33): 13,446–13,451.

Gallego-Bartolomé, J., Minguet, E.G., Marín, J.A., Prat, S., Blázquez, M.A. & Alabadí, D. (2010) Transcriptional diversification and functional conservation between DELLA proteins in *Arabidopsis*. *Molecular Biology and Evolution*, **27**(6): 1247–1256.

Grant, B.J., Rodrigues, A.P.C., ElSawy, K.M., McCammon, J.A. & Caves, L.S.D. (2006) Bio3d: an R package for the comparative analysis of protein structures. *Bioinformatics*, **22**(21): 2695–2696.

Griffiths, J., Murase, K., Rieu, I., Zentella, R., Zhang, Z.L., Powers, S.J., Gong, F., Phillips, A.L., Hedden, P., Sun, T.P. & Thomas, S.G. (2006) Genetic characterization and functional analysis of the GID1 gibberellin receptors in *Arabidopsis*. *The Plant Cell*, **18**(12): 3399–3414.

Guzzo, C.M. & Matunis, M.J. (2013) Expanding SUMO and ubiquitin-mediated signaling through hybrid SUMO-ubiquitin chains and their receptors. *Cell Cycle*, **12**(7): 1015–1017.

Hanley, J.A. & McNeil, B.J. (1982) The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, **143**(1): 29–36.

Hansen, S.K., Jamali, B. & Hubbuch, J. (2013) Selective high throughput protein quantification based on UV absorption spectra. *Biotechnology and Bioengineering*, **110**(2): 448–460.

Hao, G.F., Yang, S.G., Yang, G.F. & Zhan, C.G. (2013) Computational gibberellin-binding channel discovery unraveling the unexpected perception mechanism of hormone signal by gibberellin receptor. *Journal of Computational Chemistry*, **34**(24): 2055–2064.

Hay, R.T. (2005) SUMO: a history of modification. *Molecular Cell*, **18**(1): 1–12.

Hecker, C.M., Rabiller, M., Haglund, K., Bayer, P. & Dikic, I. (2006) Specification of SUMO1- and SUMO2-interacting motifs. *The Journal of Biological Chemistry*, **281**(23): 16,117–16,127.

Henikoff, S. & Cohen, E.H. (1984) Sequences responsible for transcription termination on a gene segment in Saccharomyces cerevisiae. *Molecular and Cellular Biology*, **4**(8): 1515–1520.

Henikoff, S. & Henikoff, J.G. (1992) Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences of the United States of America*, **89**(22): 10,915–10,919.

Hershko, A. & Ciechanover, A. (1998) The ubiquitin system. *Annual Review of Biochemistry*, **67**: 425–479.

Hirano, K., Kouketu, E., Katoh, H., Aya, K., Ueguchi-Tanaka, M. & Matsuoka, M. (2012) The suppressive function of the rice DELLA protein SLR1 is dependent on its transcriptional activation activity. *The Plant Journal*, **71**(3): 443–453.

Hoppe, A.D., Scott, B.L., Welliver, T.P., Straight, S.W. & Swanson, J.A. (2013) N-way FRET microscopy of multiple protein-protein interactions in live cells. *PloS One*, **8**(6): e64,760.

Hou, X., Lee, L.Y.C., Xia, K., Yan, Y. & Yu, H. (2010) DELLAs modulate jasmonate signaling via competitive binding to JAZs. *Developmental Cell*, **19**(6): 884–894.

Huang, L., Yang, S., Zhang, S., Liu, M., Lai, J., Qi, Y., Shi, S., Wang, J., Wang, Y., Xie, Q. & Yang, C. (2009) The *Arabidopsis* SUMO E3 ligase AtMMS21, a homologue of NSE2/MMS21, regulates cell proliferation in the root. *The Plant Journal*, **60**(4): 666–678.

Hughes, G.F. (1968) On the Mean Accuracy of Statistical Pattern Recognizers. *IEEE Transactions on Information Theory*, **8**(5): 55–63.

Hurst, R., Hook, B., Slater, M.R., Hartnett, J., Storts, D.R. & Nath, N. (2009) Protein-protein interaction studies on protein arrays: effect of detection strategies on signal-to-background ratios. *Analytical Biochemistry*, **392**(1): 45–53.

Husson, F., Josse, J., Le, S. & Mazet, J. (2013) *FactoMineR: Multivariate Exploratory Data Analysis and Data Mining with R*. R package version 1.25.

Ishida, T., Yoshimura, M., Miura, K. & Sugimoto, K. (2012) MMS21/HPY2 and SIZ1, two *Arabidopsis* SUMO E3 ligases, have distinct functions in development. *PloS One*, **7**(10): e46,897.

Iuchi, S., Suzuki, H., Kim, Y.C., Iuchi, A., Kuromori, T., Ueguchi-Tanaka, M., Asami, T., Yamaguchi, I., Matsuoka, M., Kobayashi, M. & Nakajima, M. (2007) Multiple loss-of-function of *Arabidopsis* gibberellin receptor AtGID1s completely shuts down a gibberellin signal. *The Plant Journal*, **50**(6): 958–966.

Iyer, L.M., Burroughs, A.M. & Aravind, L. (2006) The prokaryotic antecedents of the ubiquitin-signaling system and the early evolution of ubiquitin-like beta-grasp domains. *Genome Biology*, **7**(7): R60.

Jackson, S.P. & Durocher, D. (2013) Regulation of DNA damage responses by ubiquitin and SUMO. *Molecular Cell*, **49**(5): 795–807.

Jentsch, S. & Psakhye, I. (2013) Control of nuclear activities by substrate-selective and protein-group SUMOylation. *Annual Review of Genetics*, **47**: 167–186.

Kapust, R.B. & Waugh, D.S. (1999) *Escherichia coli* maltose-binding protein is uncommonly effective at promoting the solubility of polypeptides to which it is fused. *Protein Science*, **8**(8): 1668–1674.

Katari, M.S., Nowicki, S.D., Aceituno, F.F., Nero, D., Kelfer, J., Thompson, L.P., Cabello, J.M., Davidson, R.S., Goldberg, A.P., Shasha, D.E., Coruzzi, G.M. & Gutiérrez, R.A. (2010) VirtualPlant: a software platform to support systems biology research. *Plant Physiology*, **152**(2): 500–515.

Katz, C., Levy-Beladev, L., Rotem-Bamberger, S., Rito, T., Rüdiger, S.G.D. & Friedler, A. (2011) Studying protein-protein interactions using peptide arrays. *Chemical Society Reviews*, **40**(5): 2131–2145.

Kawashima, S. & Kanehisa, M. (2000) AAindex: amino acid index database. *Nucleic Acids Research*, **28**(1): 374.

Kim, J.G., Taylor, K.W., Hotson, A., Keegan, M., Schmelz, E.A. & Mudgett, M.B. (2008) XopD SUMO protease affects host transcription, promotes pathogen growth, and delays symptom development in *Xanthomonas*-infected tomato leaves. *The Plant Cell*, **20**(7): 1915–1929.

Koncz, C. & Schell, J. (1986) The promoter of T_L-DNA gene *5* controls the tissue-specific expression of chimaeric genes carried by a novel type of *Agrobacterium* binary vector. *Molecular and General Genetics*, **204**(3): 383–396.

Kroetz, M.B. & Hochstrasser, M. (2009) Identification of SUMO-interacting proteins by yeast two-hybrid analysis. *Methods in Molecular Biology*, **497**: 107–120.

Kurepa, J., Walker, J.M., Smalle, J., Gosink, M.M., Davis, S.J., Durham, T.L., Sung, D.Y. & Vierstra, R.D. (2003) The small ubiquitin-like modifier (SUMO) protein modification system in *Arabidopsis*. *The Journal of Biological Chemistry*, **278**(9): 6862–6872.

Lee, J., Nam, J., Park, H.C., Na, G., Miura, K., Jin, J.B., Yoo, C.Y., Baek, D., Kim, D.H., Jeong, J.C., Kim, D., Lee, S.Y., Salt, D.E., Mengiste, T., Gong, Q., Ma, S., Bohnert, H.J., Kwak, S.S., Bressan, R.A., Hasegawa, P.M. & Yun, D.J. (2007) Salicylic acid-mediated innate immunity in *Arabidopsis* is regulated by SIZ1 SUMO E3 ligase. *The Plant journal : for cell and molecular biology*, **49**(1): 79–90.

Lee, S., Cheng, H., King, K.E., Wang, W., He, Y., Hussain, A., Lo, J., Harberd, N.P. & Peng, J. (2002) Gibberellin regulates *Arabidopsis* seed germination via *RGL2*, a *GAI*/*RGA*-like gene whose expression is up-regulated following imbibition. *Genes & Development*, **16**(5): 646–658.

Lens, Z., Dewitte, F., Van Lint, C., de Launoit, Y., Villeret, V. & Verger, A. (2011) Purification of SUMO-1 modified IκBα and complex formation with NF-κB. *Protein Expression and Purification*, **80**(2): 211–216.

Liaw, A. & Wiener, M. (2002) Classification and Regression by randomForest. *R News*, **2**(3): 18–22.

Lin, D.Y., Huang, Y.S., Jeng, J.C., Kuo, H.Y., Chang, C.C., Chao, T.T., Ho, C.C., Chen, Y.C., Lin, T.P., Fang, H.I., Hung, C.C., Suen, C.S., Hwang, M.J., Chang, K.S., Maul, G.G. & Shih, H.M. (2006) Role of SUMO-interacting motif in Daxx SUMO modification, subnuclear localization, and repression of sumoylated transcription factors. *Molecular cell*, **24**(3): 341–354.

Ling, Y., Zhang, C., Chen, T., Hao, H., Liu, P., Bressan, R.A., Hasegawa, P.M., Jin, J.B. & Lin, J. (2012) Mutation in SUMO E3 ligase, SIZ1, disrupts the mature female gametophyte in *Arabidopsis*. *PloS One*, **7**(1): e29,470.

Liu, M., Shi, S., Zhang, S., Xu, P., Lai, J., Liu, Y., Yuan, D., Wang, Y., Du, J. & Yang, C. (2014) SUMO E3 ligase AtMMS21 is required for normal meiosis and gametophyte development in *Arabidopsis*. *BMC Plant Biology*, **14**: 153.

Locascio, A., Blázquez, M.A. & Alabadí, D. (2013) Genomic analysis of DELLA protein activity. *Plant & Cell Physiology*, **54**(8): 1229–1237.

Lois, L.M., Lima, C.D. & Chua, N.H. (2003) Small ubiquitin-like modifier modulates abscisic acid signaling in *Arabidopsis*. *The Plant Cell*, **15**(6): 1347–1359.

Ly, V., Hatherell, A., Kim, E., Chan, A., Belmonte, M.F. & Schroeder, D.F. (2013) Interactions between Arabidopsis DNA repair genes *UVH6*, *DDB1A*, and *DDB2* during abiotic stress tolerance and floral development. *Plant Science*, **213**: 88–97.

Maor, R., Jones, A., Nühse, T.S., Studholme, D.J., Peck, S.C. & Shirasu, K. (2007) Multidimensional protein identification technology (MudPIT) analysis of ubiquitinated proteins in plants. *Molecular & Cellular Proteomics*, **6**(4): 601–610.

Marblestone, J.G., Edavettal, S.C., Lim, Y., Lim, P., Zuo, X. & Butt, T.R. (2006) Comparison of SUMO fusion technology with traditional gene fusion systems: enhanced expression and solubility with SUMO. *Protein Science*, **15**(1): 182–189.

Marchler-Bauer, A., Zheng, C., Chitsaz, F., Derbyshire, M.K., Geer, L.Y., Geer, R.C., Gonzales, N.R., Gwadz, M., Hurwitz, D.I., Lanczycki, C.J., Lu, F., Lu, S., Marchler, G.H., Song, J.S., Thanki, N., Yamashita, R.A., Zhang, D. & Bryant, S.H. (2013) CDD: conserved domains and protein three-dimensional structure. *Nucleic Acids Research*, **41**(Database issue): D348–52.

Mascle, X.H., Lussier-Price, M., Cappadocia, L., Estephan, P., Raiola, L., Omichinski, J.G. & Aubry, M. (2013) Identification of a non-covalent ternary complex formed by PIAS1, SUMO1, and UBC9 proteins involved in transcriptional regulation. *The Journal of Biological Chemistry*, **288**(51): 36,312–36,327.

Mazur, M.J. & van den Burg, H.A. (2012) Global SUMO Proteome Responses Guide Gene Regulation, mRNA Biogenesis, and Plant Stress Responses. *Frontiers in Plant Science*, **3**: 215.

Middleton, A.M., Úbeda-Tomás, S., Griffiths, J., Holman, T., Hedden, P., Thomas, S.G., Phillips, A.L., Holdsworth, M.J., Bennett, M.J., King, J.R. & Owen, M.R. (2012) Mathematical modeling elucidates the role of transcriptional feedback in gibberellin signaling. *Proceedings of the National Academy of Sciences of the United States of America*, **109**(19): 7571–7576.

Miller, M.J., Barrett-Wilt, G.A., Hua, Z. & Vierstra, R.D. (2010) Proteomic analyses identify a diverse array of nuclear processes affected by small ubiquitin-like modifier conjugation in *Arabidopsis*. *Proceedings of the National Academy of Sciences of the United States of America*, **107**(38): 16,512–16,517.

Miller, M.J., Scalf, M., Rytz, T.C., Hubler, S.L., Smith, L.M. & Vierstra, R.D. (2012) Quantitative proteomics reveal factors regulating RNA biology as dynamics targets of stress-induced sumoylation in *Arabidopsis*. *Molecular & cellular proteomics*, **12**(2): 449–463.

Minty, A., Dumont, X., Kaghad, M. & Caput, D. (2000) Covalent modification of p73alpha by SUMO-1. Two-hybrid screening with p73 identifies novel SUMO-1-interacting proteins and a SUMO-1 interaction motif. *The Journal of Biological Chemistry*, **275**(46): 36,316–36,323.

Miura, K. & Hasegawa, P.M. (2010) Sumoylation and other ubiquitin-like post-translational modifications in plants. *Trends in Cell Biology*, **20**(4): 223–232.

Miura, K., Jin, J.B. & Hasegawa, P.M. (2007) Sumoylation, a post-translational regulatory process in plants. *Current Opinion in Plant Biology*, **10**(5): 495–502.

Miura, K., Lee, J., Gong, Q., Ma, S., Jin, J.B., Yoo, C.Y., Miura, T., Sato, A., Bohnert, H.J. & Hasegawa, P.M. (2011) *SIZ1* regulation of phosphate starvation-induced root architecture remodeling involves the control of auxin accumulation. *Plant Physiology*, **155**(2): 1000–1012.

Miura, K., Lee, J., Jin, J.B., Yoo, C.Y., Miura, T. & Hasegawa, P.M. (2009) Sumoylation of ABI5 by the Arabidopsis SUMO E3 ligase SIZ1 negatively regulates abscisic acid signaling. *Proceedings of the National Academy of Sciences of the United States of America*, **106**(13): 5418–5423.

Miura, K., Lee, J., Miura, T. & Hasegawa, P.M. (2010) SIZ1 controls cell growth and plant development in *Arabidopsis* through salicylic acid. *Plant & Cell Physiology*, **51**(1): 103–113.

Miura, K. & Nozawa, R. (2014) Overexpression of *SIZ1* enhances tolerance to cold and salt stresses and attenuates response to abscisic acid in *Arabidopsis thaliana*. *Plant Biotechnology*.

Miura, K., Okamoto, H., Okuma, E., Shiba, H., Kamada, H., Hasegawa, P.M. & Murata, Y. (2012) *SIZ1* deficiency causes reduced stomatal aperture and enhanced drought tolerance via controlling salicylic acid-induced accumulation of reactive oxygen species in *Arabidopsis*. *The Plant Journal*.

Miura, K., Rus, A., Sharkhuu, A., Yokoi, S., Karthikeyan, A.S., Raghothama, K.G., Baek, D., Koo, Y.D., Jin, J.B., Bressan, R.A., Yun, D.J. & Hasegawa, P.M. (2005) The *Arabidopsis* SUMO E3 ligase SIZ1 controls phosphate deficiency responses. *Proceedings of the National Academy of Sciences of the United States of America*, **102**(21): 7760–7765.

Mohideen, F., Capili, A.D., Bilimoria, P.M., Yamada, T., Bonni, A. & Lima, C.D. (2009) A molecular basis for phosphorylation-dependent SUMO conjugation by the E2 UBC9. *Nature Structural & Molecular Biology*, **16**(9): 945–952.

Morris, J.R., Boutell, C., Keppler, M., Densham, R., Weekes, D., Alamshah, A., Butler, L., Galanty, Y., Pangon, L., Kiuchi, T., Ng, T. & Solomon, E. (2009) The SUMO modification pathway is involved in the BRCA1 response to genotoxic stress. *Nature*, **462**(7275): 886–890.

Murase, K., Hirano, Y., Sun, T.P. & Hakoshima, T. (2008) Gibberellin-induced DELLA recognition by the gibberellin receptor GID1. *Nature*, **456**(7221): 459–463.

Nakajima, M., Shimada, A., Takashi, Y., Kim, Y.C., Park, S.H., Ueguchi-Tanaka, M., Suzuki, H., Katoh, E., Iuchi, S., Kobayashi, M., Maeda, T., Matsuoka, M. & Yamaguchi, I. (2006) Identification and characterization of *Arabidopsis* gibberellin receptors. *The Plant Journal*, **46**(5): 880–889.

Namanja, A.T., Li, Y.J., Su, Y., Wong, S., Lu, J., Colson, L.T., Wu, C., Li, S.S.C. & Chen, Y. (2012) Insights into high affinity small ubiquitin-like modifier (SUMO) recognition by SUMO-interacting motifs (SIMs) revealed by a combination of NMR and peptide array analysis. *The Journal of Biological Chemistry*, **287**(5): 3231–3240.

Nelis, S. (2011) *Sumoylation of the DELLA proteins REPRESSOR OF ga1-3 and GA INSENSITIVE in* Arabidopsis thaliana *offers a new perspective on gibberellin mediated regulation of growth in plants*. Master's thesis, University of Warwick, United Kingdom.

Novatchkova, M., Tomanov, K., Hofmann, K., Stuible, H.P. & Bachmair, A. (2012) Update on sumoylation: defining core components of the plant SUMO conjugation system by phylogenetic comparison. *The New Phytologist*, **195**(1): 23–31.

Ogawa, M., Hanada, A., Yamauchi, Y., Kuwahara, A., Kamiya, Y. & Yamaguchi, S. (2003) Gibberellin biosynthesis and response during *Arabidopsis* seed germination. *The Plant Cell*, **15**(7): 1591–1604.

Ogrocká, A., Polanská, P., Majerová, E., Janeba, Z., Fajkus, J. & Fojtová, M. (2014) Compromised telomere maintenance in hypomethylated *Arabidopsis thaliana* plants. *Nucleic Acids Research*, **42**(5): 2919–2931.

Okada, S., Nagabuchi, M., Takamura, Y., Nakagawa, T., Shinmyozu, K., Nakayama, J.i. & Tanaka, K. (2009) Reconstitution of *Arabidopsis thaliana* SUMO pathways in *E. coli*: functional evaluation of SUMO machinery proteins and mapping of SUMOylation sites by mass spectrometry. *Plant & Cell Physiology*, **50**(6): 1049–1061.

Ouyang, K.J., Woo, L.L., Zhu, J., Huo, D., Matunis, M.J. & Ellis, N.A. (2009) SUMO modification regulates BLM and RAD51 interaction at damaged replication forks. *PLoS Biology*, **7**(12): e1000,252.

Park, B.S., Song, J.T. & Seo, H.S. (2011) *Arabidopsis* nitrate reductase activity is stimulated by the E3 SUMO ligase AtSIZ1. *Nature Communications*, **2**: 400.

Park-Sarge, O.K. & Sarge, K.D. (2009) Detection of Sumoylated Proteins. In *The Protein Protocols Handbook* (edited by J.M. Walker). Humana Press, Hatfield.

Parry, G., Calderon-Villalobos, L.I., Prigge, M., Peret, B., Dharmasiri, S., Itoh, H., Lechner, E., Gray, W.M., Bennett, M. & Estelle, M. (2009) Complex regulation of the TIR1/AFB family of auxin receptors. *Proceedings of the National Academy of Sciences of the United States of America*, **106**(52): 22,540–22,545.

Pei, J. & Grishin, N.V. (2001) AL2CO: calculation of positional conservation in a protein sequence alignment. *Bioinformatics*, **17**(8): 700–712.

Percherancier, Y., Germain-Desprez, D., Galisson, F., Mascle, X.H., Dianoux, L., Estephan, P., Chelbi-Alix, M.K. & Aubry, M. (2009) Role of SUMO in RNF4-mediated promyelocytic leukemia protein (PML) degradation: sumoylation of PML and phospho-switch control of its SUMO binding domain dissected in living cells. *The Journal of Biological Chemistry*, **284**(24): 16,595–16,608.

Petrásek, J., Mravec, J., Bouchard, R., Blakeslee, J.J., Abas, M., Seifertová, D., Wisniewska, J., Tadele, Z., Kubes, M., Covanová, M., Dhonukshe, P., Skupa, P., Benková, E., Perry, L., Krecek, P., Lee, O.R., Fink, G.R., Geisler, M., Murphy, A.S., Luschnig, C., Zazímalová, E. & Friml, J. (2006) PIN proteins perform a rate-limiting function in cellular auxin efflux. *Science*, **312**(5775): 914–918.

Prlić, A., Domingues, F.S. & Sippl, M.J. (2000) Structure-derived substitution matrices for alignment of distantly related sequences. *Protein Engineering*, **13**(8): 545–550.

Qin, Q., Wang, W., Guo, X., Yue, J., Huang, Y., Xu, X., Li, J. & Hou, S. (2014) *Arabidopsis* DELLA protein degradation is controlled by a type-one protein phosphatase, TOPP4. *PLoS Genetics*, **10**(7): e1004,464.

Qüesta, J.I., Fina, J.P. & Casati, P. (2013) DDM1 and ROS1 have a role in UV-B induced- and oxidative DNA damage in *A. thaliana*. *Frontiers in Plant Science*, **4**: 420.

R Core Team (2013) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Reeves, P.H., Murtas, G., Dash, S. & Coupland, G. (2002) *early in short days 4*, a mutation in *Arabidopsis* that causes early flowering and reduces the mRNA abundance of the floral repressor *FLC*. *Development*, **129**(23): 5349–5361.

Ren, J., Gao, X., Jin, C., Zhu, M., Wang, X., Shaw, A., Wen, L., Yao, X. & Xue, Y. (2009) Systematic study of protein sumoylation: Development of a site-specific predictor of SUMOsp 2.0. *Proteomics*, **9**(12): 3409–3412.

Reverter, D. & Lima, C.D. (2006) Structural basis for SENP2 protease interactions with SUMO precursors and conjugated substrates. *Nature structural & Molecular Biology*, **13**(12): 1060–1068.

Roumeliotis, E., Visser, R.G.F. & Bachem, C.W.B. (2012) A crosstalk of auxin and GA during tuber development. *Plant Signaling & Behavior*, **7**(10): 1360–1363.

RStudio (2014) *shiny: Web Application Framework for R*. R package version 0.10.0.

Sadanandom, A., Bailey, M., Ewan, R., Lee, J. & Nelis, S. (2012) The ubiquitin-proteasome system: central modifier of plant signalling. *The New Phytologist*, **196**(1): 13–28.

Saracco, S.A., Miller, M.J., Kurepa, J. & Vierstra, R.D. (2007) Genetic analysis of SUMOylation in Arabidopsis: conjugation of SUMO1 and SUMO2 to nuclear proteins is essential. *Plant Physiology*, **145**(1): 119–134.

Schindelin, J., Arganda-Carreras, I., Frise, E., Kaynig, V., Longair, M., Pietzsch, T., Preibisch, S., Rueden, C., Saalfeld, S., Schmid, B., Tinevez, J.Y., White, D.J., Hartenstein, V., Eliceiri, K., Tomancak, P. & Cardona, A. (2012) Fiji: an open-source platform for biological-image analysis. *Nature Methods*, **9**(7): 676–682.

Schneider, T.D. & Stephens, R.M. (1990) Sequence logos: a new way to display consensus sequences. *Nucleic Acids Research*, **18**(20): 6097–6100.

Schwamborn, K., Knipscheer, P., van Dijk, E., van Dijk, W.J., Sixma, T.K., Meloen, R.H. & Langedijk, J.P.M. (2008) SUMO assay with peptide arrays on solid support: insights into SUMO target sites. *Journal of Biochemistry*, **144**(1): 39–49.

Schwechheimer, C. & Willige, B.C. (2009) Shedding light on gibberellic acid signalling. *Current Opinion in Plant Biology*, **12**(1): 57–62.

Sekiyama, N., Ikegami, T., Yamane, T., Ikeguchi, M., Uchimura, Y., Baba, D., Ariyoshi, M., Tochio, H., Saitoh, H. & Shirakawa, M. (2008) Structure of the small ubiquitin-like modifier (SUMO)-interacting motif of MBD1-containing chromatin-associated factor 1 bound to SUMO-3. *The Journal of Biological Chemistry*, **283**(51): 35,966–35,975.

Shaked, H., Avivi-Ragolsky, N. & Levy, A.A. (2006) Involvement of the Arabidopsis *SWI2/SNF2* chromatin remodeling gene family in DNA damage response and recombination. *Genetics*, **173**(2): 985–994.

Shimada, A., Ueguchi-Tanaka, M., Nakatsu, T., Nakajima, M., Naoe, Y., Ohmiya, H., Kato, H. & Matsuoka, M. (2008) Structural basis for gibberellin recognition by its receptor GID1. *Nature*, **456**(7221): 520–523.

Shultzaberger, R.K., Bucheimer, R.E., Rudd, K.E. & Schneider, T.D. (2001) Anatomy of *Escherichia coli* ribosome binding sites. *Journal of Molecular Biology*, **313**(1): 215–228.

Sing, T., Sander, O., Beerenwinkel, N. & Lengauer, T. (2005) ROCR: visualizing classifier performance in R. *Bioinformatics*, **21**(20): 7881.

Son, G.H., Park, B.S., Song, J.T. & Seo, H.S. (2014) FLC-mediated flowering repression is positively regulated by sumoylation. *Journal of Experimental Botany*, **65**(1): 339–351.

Song, J., Durrin, L.K., Wilkinson, T.A., Krontiris, T.G. & Chen, Y. (2004) Identification of a SUMO-binding motif that recognizes SUMO-modified proteins. *Proceedings of the National Academy of Sciences of the United States of America*, **101**(40): 14,373–14,378.

Song, J., Zhang, Z., Hu, W. & Chen, Y. (2005) Small ubiquitin-like modifier (SUMO) recognition of a SUMO binding motif: a reversal of the bound orientation. *The Journal of Biological Chemistry*, **280**(48): 40,122–40,129.

Stehmeier, P. & Muller, S. (2009) Phospho-regulated SUMO interaction modules connect the SUMO system to CK2 signaling. *Molecular Cell*, **33**(3): 400–409.

Sun, T.P. & Gubler, F. (2004) Molecular mechanism of gibberellin signaling in plants. *Annual Review of Plant Biology*, **55**: 197–223.

Sun, X., Jones, W.T., Harvey, D., Edwards, P.J.B., Pascal, S.M., Kirk, C., Considine, T., Sheerin, D.J., Rakonjac, J., Oldfield, C.J., Xue, B., Dunker, A.K. & Uversky, V.N. (2010) N-terminal domains of DELLA proteins are intrinsically unstructured in the absence of interaction with GID1/gibberellic acid receptors. *The Journal of Biological Chemistry*, **285**(15): 11,557–11,571.

Suzuki, H., Park, S.H., Okubo, K., Kitamura, J., Ueguchi-Tanaka, M., Iuchi, S., Katoh, E., Kobayashi, M., Yamaguchi, I., Matsuoka, M., Asami, T. & Nakajima, M. (2009) Differential expression and affinities of *Arabidopsis* gibberellin receptors can explain variation in phenotypes of multiple knock-out mutants. *The Plant Journal*, **60**(1): 48–55.

Tatham, M.H., Geoffroy, M.C., Shen, L., Plechanovová, A., Hattersley, N., Jaffray, E.G., Palvimo, J.J. & Hay, R.T. (2008) RNF4 is a poly-SUMO-specific E3 ubiquitin ligase required for arsenic-induced PML degradation. *Nature Cell Biology*, **10**(5): 538–546.

Tatham, M.H., Kim, S., Jaffray, E., Song, J., Chen, Y. & Hay, R.T. (2005) Unique binding interactions among Ubc9, SUMO and RanBP2 reveal a mechanism for SUMO paralog selection. *Nature Structural & Molecular Biology*, **12**(1): 67–74.

Teng, S., Luo, H. & Wang, L. (2012) Predicting protein sumoylation sites from sequence features. *Amino Acids*, **43**(1): 447–455.

Tinevez, J.Y. (2013) *Circle pixel coordinates using mid-point algorithm*. MATLAB Central File Exchange, url: http://www.mathworks.co.uk/matlabcentral/fileexchange/33844-circle-pixel-coordinates-using-mid-point-algorithm/.

Truong, K., Su, Y., Song, J. & Chen, Y. (2011) Entropy-driven mechanism of an E3 ligase. *Biochemistry*, **50**(25): 5757–5766.

Tyler, L., Thomas, S.G., Hu, J., Dill, A., Alonso, J.M., Ecker, J.R. & Sun, T.P. (2004) DELLA proteins and gibberellin-regulated seed germination and floral development in *Arabidopsis*. *Plant Physiology*, **135**(2): 1008–1019.

Ueguchi-Tanaka, M., Ashikari, M., Nakajima, M., Itoh, H., Katoh, E., Kobayashi, M., Chow, T.y., Hsing, Y.i.C., Kitano, H., Yamaguchi, I. & Matsuoka, M. (2005) *GIBBERELLIN INSENSITIVE DWARF1* encodes a soluble receptor for gibberellin. *Nature*, **437**(7059): 693–698.

Ueguchi-Tanaka, M., Nakajima, M., Katoh, E., Ohmiya, H., Asano, K., Saji, S., Hongyu, X., Ashikari, M., Kitano, H., Yamaguchi, I. & Matsuoka, M. (2007) Molecular interactions of a soluble gibberellin receptor, GID1, with a rice DELLA protein, SLR1, and gibberellin. *The Plant Cell*, **19**(7): 2140–2155.

Ullmann, R., Chien, C.D., Avantaggiati, M.L. & Muller, S. (2012) An acetylation switch regulates SUMO-dependent protein interaction networks. *Molecular Cell*, **46**(6): 759–770.

Ulrich, H.D. (2005) SUMO modification: wrestling with protein conformation. *Current Biology*, **15**(7): R257–9.

Ulrich, H.D. (2008) The fast-growing business of SUMO chains. *Molecular Cell*, **32**(3): 301–305.

van den Burg, H.A., Kini, R.K., Schuurink, R.C. & Takken, F.L.W. (2010) *Arabidopsis* small ubiquitin-like modifier paralogs have distinct functions in development and defense. *The Plant Cell*, **22**(6): 1998–2016.

Vincentelli, R., Canaan, S., Offant, J., Cambillau, C. & Bignon, C. (2005) Automated expression and solubility screening of His-tagged proteins in 96-well format. *Analytical Biochemistry*, **346**(1): 77–84.

Vogt, B. & Hofmann, K. (2012) Bioinformatical detection of recognition factors for ubiquitin and SUMO. *Methods in Molecular Biology*, **832**: 249–261.

Wallrabe, H., Sun, Y., Fang, X., Periasamy, A. & Bloom, G. (2012) Three-Color FRET expands the ability to quantify the interactions of several proteins involved in actin filament nucleation. *Proceedings of SPIE*, **8226**.

Wang, F., Zhu, D., Huang, X., Li, S., Gong, Y., Yao, Q., Fu, X., Fan, L.M. & Deng, X.W. (2009) Biochemical insights on degradation of *Arabidopsis* DELLA proteins gained from a cell-free assay system. *The Plant Cell*, **21**(8): 2378–2390.

Weber, A.R., Schuermann, D. & Schär, P. (2014) Versatile recombinant SUMOylation system for the production of SUMO-modified protein. *PloS One*, **9**(7): e102,157.

Weiner, M.P., Costa, G.L., Schoettlin, W., Cline, J., Mathur, E. & Bauer, J.C. (1994) Site-directed mutagenesis of double-stranded DNA by the polymerase chain reaction. *Gene*, **151**(1-2): 119–123.

Weiser, A.A., Or-Guil, M., Tapia, V., Leichsenring, A., Schuchhardt, J., Frömmel, C. & Volkmer-Engert, R. (2005) SPOT synthesis: reliability of array-based measurement of peptide binding affinity. *Analytical Biochemistry*, **342**(2): 300–311.

Wickham, H. (2009) *ggplot2: elegant graphics for data analysis*. Springer New York. ISBN 978-0-387-98140-6.

Wilkinson, K.A. & Henley, J.M. (2010) Mechanisms, regulation and consequences of protein SUMOylation. *The Biochemical Journal*, **428**(2): 133–145.

Woodcock, A. (2014) *The effect of SUMOylation on DELLA proteins and abiotic stress responses in* Arabidopsis thaliana. Ph.D. thesis, University of Warwick, United Kingdom.

Xu, P., Yuan, D., Liu, M., Li, C., Liu, Y., Zhang, S., Yao, N. & Yang, C. (2013) AtMMS21, an SMC5/6 complex subunit, is involved in stem cell niche maintenance and DNA damage responses in *Arabidopsis* roots. *Plant Physiology*, **161**(4): 1755–1768.

Xue, Y., Zhou, F., Fu, C., Xu, Y. & Yao, X. (2006) SUMOsp: a web server for sumoylation site prediction. *Nucleic Acids Research*, **34**(Web Server issue): 254–7.

Yamamoto, Y., Hirai, T., Yamamoto, E., Kawamura, M., Sato, T., Kitano, H., Matsuoka, M. & Ueguchi-Tanaka, M. (2010) A rice gid1 suppressor mutant reveals that gibberellin is not always required for interaction between its receptor, GID1, and DELLA proteins. *The Plant Cell*, **22**(11): 3589–3602.

Yang, S.H., Galanis, A., Witty, J. & Sharrocks, A.D. (2006) An extended consensus motif enhances the specificity of substrate modification by SUMO. *The EMBO Journal*, **25**(21): 5083–5093.

Yang, S.H. & Sharrocks, A.D. (2010) The SUMO E3 ligase activity of Pc2 is coordinated through a SUMO interaction motif. *Molecular and cellular biology*, **30**(9): 2193–2205.

Yoo, C.Y., Miura, K., Jin, J.B., Lee, J., Park, H.C., Salt, D.E., Yun, D.J., Bressan, R.A. & Hasegawa, P.M. (2006) SIZ1 small ubiquitin-like modifier E3 ligase facilitates basal thermotolerance in *Arabidopsis* independent of salicylic acid. *Plant Physiology*, **142**(4): 1548–1558.

Zhang, S., Qi, Y., Liu, M. & Yang, C. (2013) SUMO E3 ligase AtMMS21 regulates drought tolerance in *Arabidopsis thaliana*. *Journal of Integrative Plant Biology*, **55**(1): 83–95.

Zheng, Y., Schumaker, K.S. & Guo, Y. (2012) Sumoylation of transcription factor MYB30 by the small ubiquitin-like modifier E3 ligase SIZ1 mediates abscisic acid response in *Arabidopsis thaliana*. *Proceedings of the National Academy of Sciences of the United States of America*.

Zhu, S., Goeres, J., Sixt, K.M., Békés, M., Zhang, X.D., Salvesen, G.S. & Matunis, M.J. (2009) Protection from isopeptidase-mediated deconjugation regulates paralog-selective sumoylation of RanGAP1. *Molecular Cell*, **33**(5): 570–580.

Zuo, X., Li, S., Hall, J., Mattern, M.R., Tran, H., Shoo, J., Tan, R., Weiss, S.R. & Butt, T.R. (2005) Enhanced expression and purification of membrane proteins by SUMO fusion in Escherichia coli. *Journal of Structural and Functional Genomics*, **6**(2-3): 103–111.

Appendix A

Molecular Biology

A.1 DNA primers for PCR

Name	Sequence (5' - 3')	T_a (°C)
D_AttB1-F	ACAAGTTTGTACAAAAAAGCAGGCT	55
D_AttB2-R	ACCACTTTGTACAAGAAAGCTGGGT	55
E_RGA-F	CAGGCCCAGCCGGCCATGAAGAGAGATCATCACCAATTCCAAG	58
E_SUM1-F	CAGGCCCAGCCGGCCATGTCTGCAAACCAGGAGGAAGAC	58
E_RGA-R	GCTACTTCTAGAATGTACGCCGCCGTCGAGAGTTTC	58
C_RGA-F	CACCATGAAGAGAGATCATCACCAATT	58
C_RGA-R	GTACGCCGCCGTCGAGAGTTT	58
C_SCE1-F	CACCATGGCTAGTGGAATCGCT	60
C_SCE1-R	TTAGACAAGAGCAGGATACTGC	60
M13 (-20)-F	GTAAAACGACGGCCAG	55
M13-R	CAGGAAACAGCTATGAC	55
M_GID1a_V22A-F	GTGGTTCCTCTCAATACATGGGCTTTAATATCCAACTTCAAAGTA	55
M_GID1a_V22A-R	TACTTTGAAGTTGGATATTAAAGCCCATGTATTGAGAGGAACCAC	55
M_GID1a_V22S-F	ACAGTGGTTCCTCTCAATACATGGAGTTTAATATCCAACTTCAAAGTAGC	55
M_GID1a_V22S-R	GCTACTTTGAAGTTGGATATTAAACTCCATGTATTGAGAGGAACCACTGT	55
D_OTS1-F	CACCATGACGAAGAGGAAGAAGGA	60
D_OTS1-R	TTACTCTGTCTGGTCACTGACAC	60
D_OTS2-F	CACCATGAAGAGACAAAGAGCAATCG	60
D_OTS2-R	TTAATCTGTTTGGTTACCCTTGC	60
D_T-DNA-R	TGGTTCACGTAGTGGGCCATCG	60
D_ots1-F	CGACAAGAAGTGGTTTAGACC	60
D_ots2-F	GACAGGGATGCATATTTTGTGAAG	60

Table A.1: DNA primers used for PCR.

A.2 Genotyping of *ots1 ots2* T-DNA insertion lines

T-DNA insertion knock-down of *ots1-1 ots2-1* were confirmed by PCR. Primers for the full length genes were used to confirm absence of the functional *OTS1* and *OTS2* genes. T-DNA insertions were confirmed using a left border primer for the T-DNA insertion plasmid pROK 2 and a gene specific primer for *OTS1* or *OTS2*. The line *ots1-1 ots2-1* displayed the expected lack of *OTS1* or *OTS2* fragments and the expected T-DNA insert fragments for *ots1-1* and *ots2-1* (see Figure A.1).

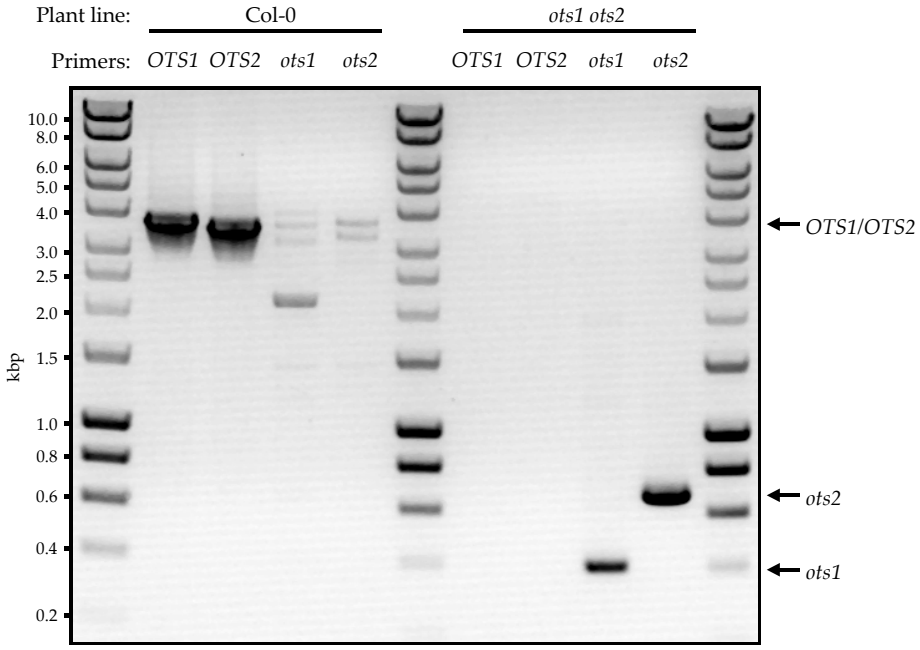


Figure A.1: Confirmation of the *ots1-1 ots2-1* knock-down plant line.

Appendix B

SIM peptide arrays

B.1 Areas of images sampled for baseline subtraction assessment

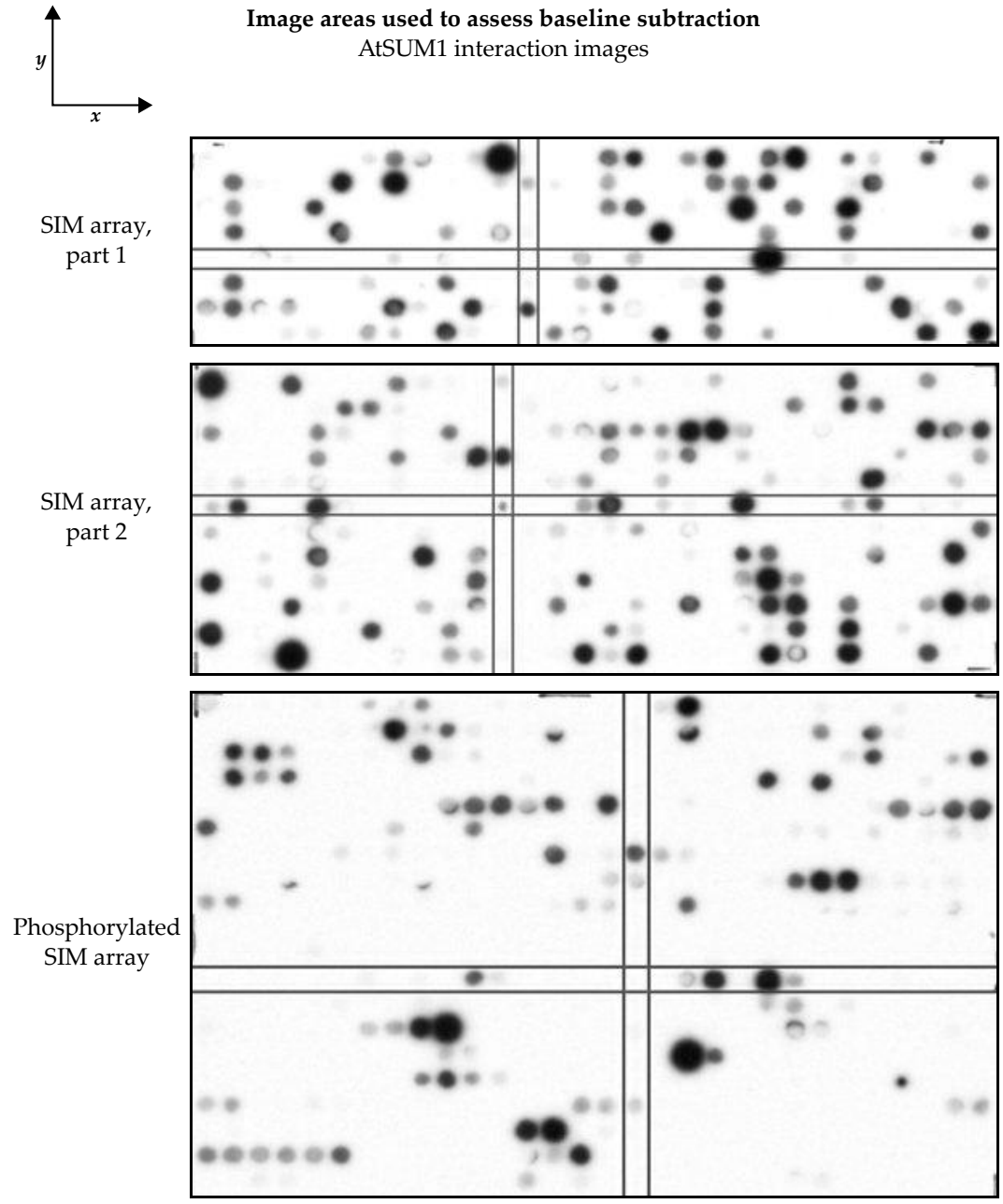


Figure B.1: Peptide array image areas used to assess baseline flattening for the AtSUM1 interaction far-western blot.

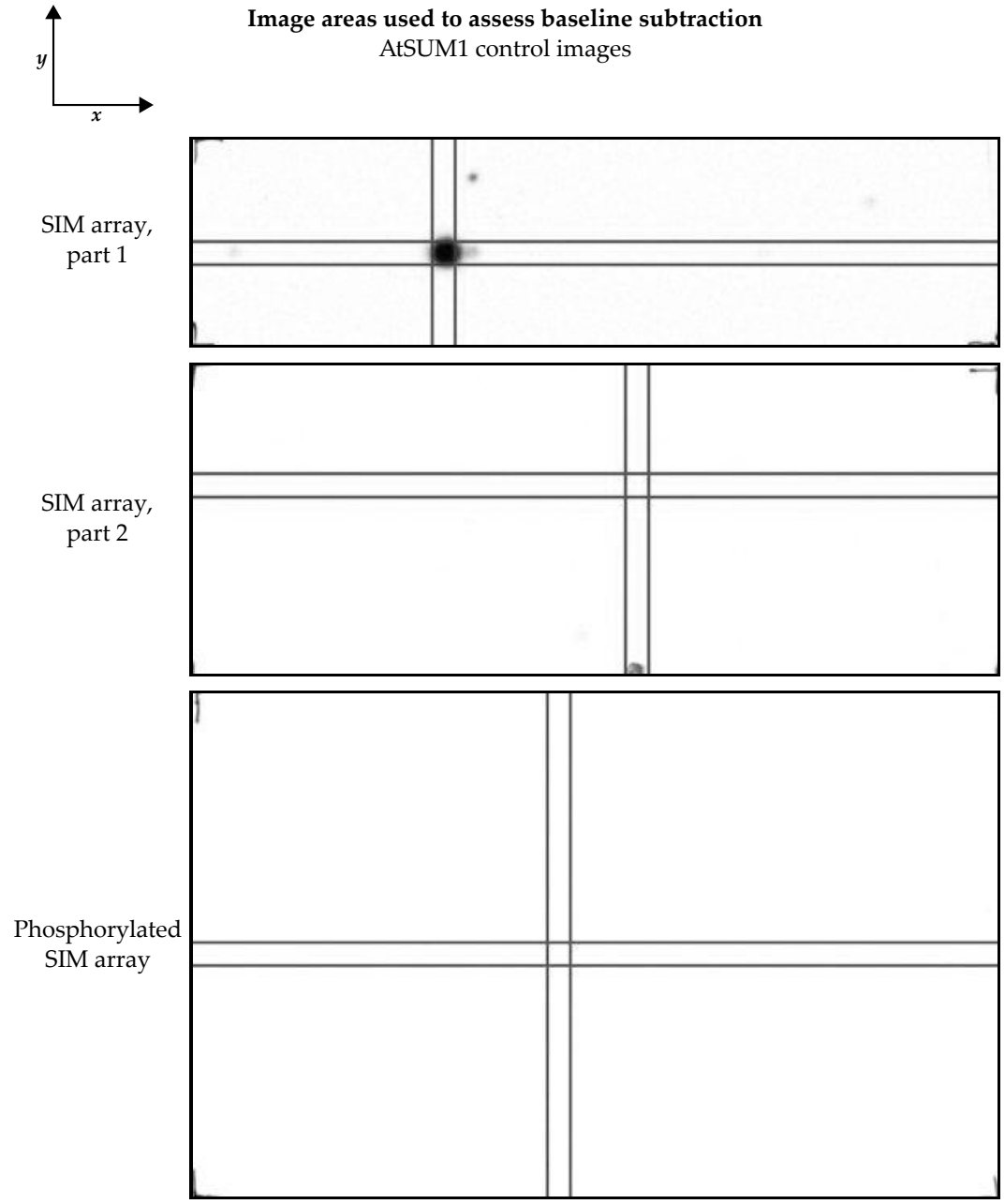


Figure B.2: Peptide array image areas used to assess baseline flattening for the AtSUM1 control far-western blot.

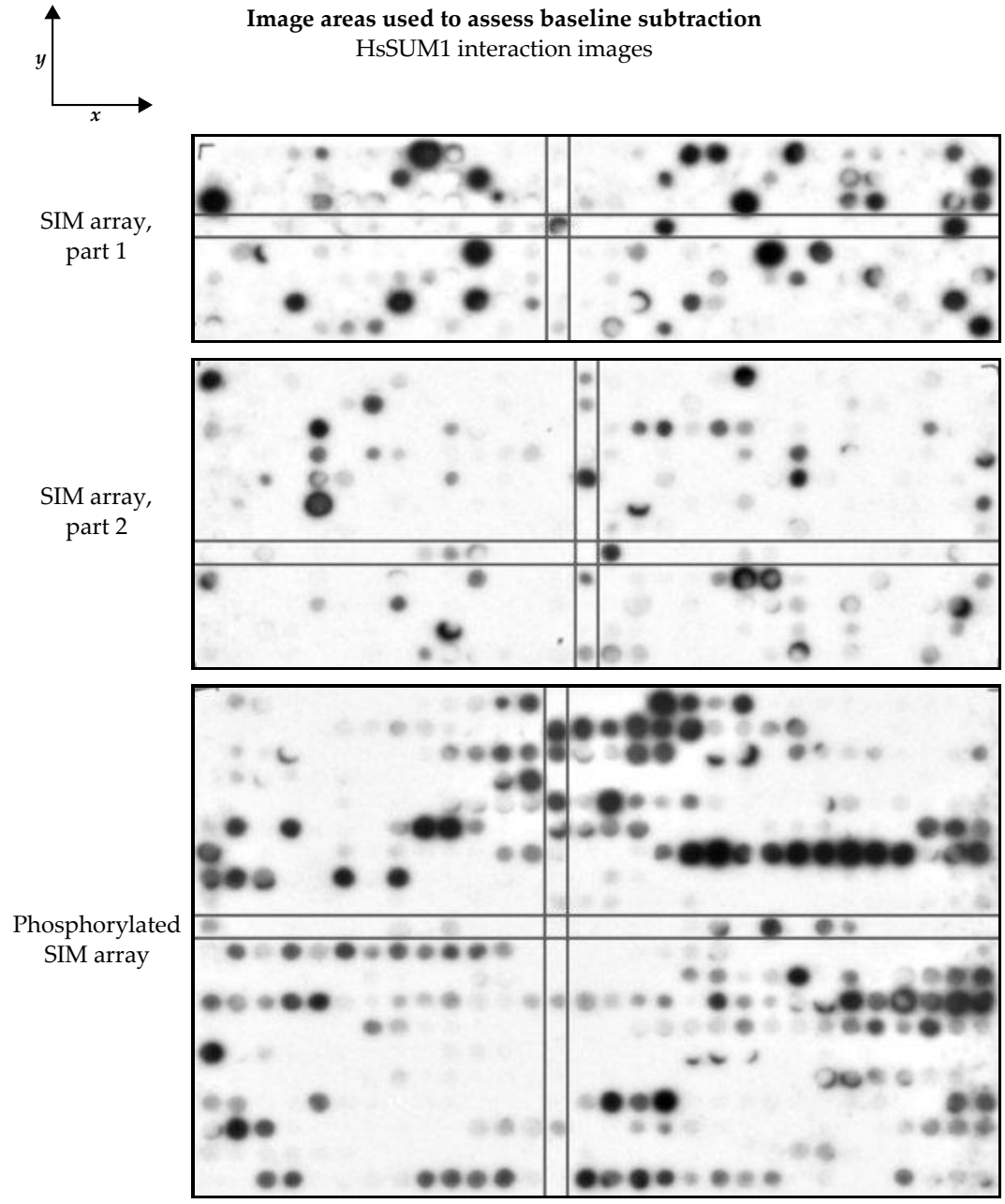


Figure B.3: Peptide array image areas used to assess baseline flattening for the HsSUM1 interaction far-western blot.

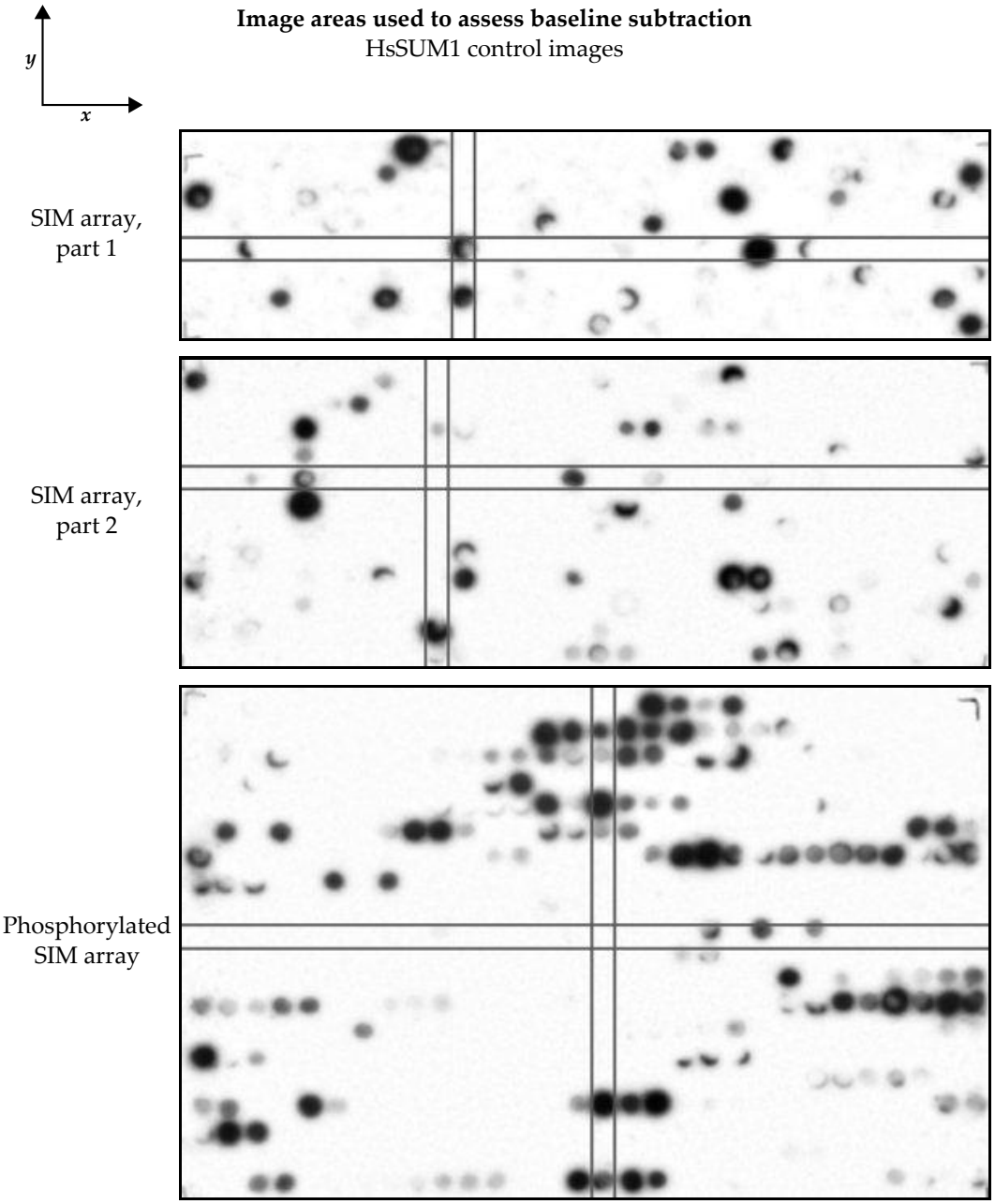


Figure B.4: Peptide array image areas used to assess baseline flattening for the HsSUM1 control far-western blot.

B.2 SIM peptide sequences

Table B.1: SIM peptide interaction values. NA indicates data that was removed due to quality issues. A lowercase 'p' before an amino acid indicates a phosphorylated amino acid. See Table 3.1 for explanation of peptide classes. Peptide classes not in this table are control peptides from known SIMs.

Peptide	Class	AtSUM1	HsSUM1
FEILDLTIGEWRT	B4	0.03	-0.03
GKPVDLTPEQEEV	B1	0.00	-0.01
GWIWGWEEEAIDV	A3	NA	NA
SWTYDLTSDPLEN	B1	NA	NA
CSLSESELDLLIS	R1	NA	0.39
PIMPTVSSSFDLK	A3	0.00	-0.01
YLLVNIGSGVSMI	A1	NA	NA
LRVVDLTILGFLF	B1	0.51	NA
KKLMDIKNRVTSR	A1	0.18	NA
DMIILVSKHVTFK	A1	0.01	NA
SKVINLSEDIFAG	A2	0.16	0.00
DLVLDLCKQMELK	B4	1.03	0.00
GDKQDESFEPFIP	R3	0.05	-0.02
PLPPSPSDSFNSV	A3	NA	NA
EAIVSVSSDSEIP	A2	NA	NA
SVKGSSSLPLWFS	R3	0.59	-0.01
KPLLFLSSSSSFS	A2	0.80	0.08
DFSDESDLNMLWS	R2	NA	NA
QLVIDLTRNKGMR	B1	0.37	NA
RSLMMLSRDSSFK	A1	0.92	NA
PMMIVVECNPKGD	A1	0.00	0.01
AMKNTSHVLLLSL	R1	0.67	0.03
GFSPDLTTFNIRA	B2	0.99	NA
LILPDLTSGAESK	B2	0.00	0.00
LNYVRLSESSFAS	A3	0.44	0.16
KSILYLNRHQTEE	A1	0.13	0.14
PLPFNYDEDDEKD	A3	0.00	0.02
RKLMFLSDDGIMF	A2	0.55	0.06
YHYIDLSKNEISG	B3	0.00	0.65
ELELDLSSNAKVQ	B4	0.00	0.03
KPLVIVCNKTDLM	A1	0.00	NA
LIESRTVVPLNTW	GID1a-SIM-A	0.57	-0.01
SFRPDLYSWPSLP	B2	NA	NA
TDAESEELRVMRE	R3	0.00	-0.02
VCLLDLEYGLPVQ	B4	NA	NA
SVAIDLCGMTQAE	B3	0.92	NA
IQSAEGSVDVLMD	R1	NA	NA
QNPISLSSSVSFQ	A2	0.99	NA
RHMIPVSEENMYP	A1	0.00	NA
TCIVDLYTHLPKV	B1	0.00	0.01
ADYIDLTDLTVHD	B1	NA	0.91
NKVVIMSSSHSSI	A3	0.16	0.00
VNHIDLYRLDFQT	B1	0.20	0.01
CAVPGCDVKVMSD	R1	0.04	0.00
DDIIKLDEEALKA	A2	0.00	-0.02
EFKLDLTCGINPS	B2	0.29	0.00
DPSVDLSERPMMY	B3	NA	NA
SHCVDLEMLDEVK	B4	0.00	0.56
KMIVDLEPPRFDN	B3	0.00	0.00
IAYWRNHVVVVIL	R1	0.56	0.02
HIPIVVGSQQRYE	A1	0.52	0.01
RVIIDLDFKTQFE	B3	0.74	0.25
ETQDSSSLTPFYD	R3	NA	NA
KWMDTRDVLIVNP	R1	0.00	-0.03

Peptide	Class	AtSUM1	HsSUM1
DGFLLVEESLRIK	A3	0.06	0.37
DSDKESSLNLILR	R2	0.68	NA
SDMIAMTNEDLNK	A1	0.00	-0.03
VAEEESELCMLME	R2	NA	NA
EEMKESSISMVEA	R2	0.00	-0.04
PVPFDLTPRLSNK	B2	0.42	NA
CEIFDLDSEHVRI	A3	0.00	NA
IVPLNTWVLISNF	GID1a-SIM-B	0.39	NA
ESIVFMEDDEVKE	A3	0.00	-0.04
PPAVDLELEKVLG	B3	0.00	-0.03
VTLLRVQDSQNRS	A1	0.71	0.44
KKEFETQMEVVGK	R1	-0.01	0.01
LQSFDLSSNKFNG	B4	0.00	0.06
QGLMFVGESALSL	A1	NA	NA
DRILGVSDEKKPI	A2	0.00	-0.02
LCLLVMKNSQVSK	A1	0.00	-0.02
WSDQDSSLWMIPC	R2	NA	NA
LLLVSLCDHKTLS	A1	0.00	0.14
NDVLQISDHLSKQ	A1	0.00	-0.01
GEEVDLYMDQSSG	B2	NA	NA
GILIDLEPARTED	B3	NA	NA
TNPTSSSFSFPLT	R3	0.47	NA
ATVVDLYRTRYKY	B1	0.69	-0.03
RDMIVVGESGLFT	A1	NA	NA
TQDTSSSVLLFFD	R3	NA	NA
TGMLKMEEEFVDS	A2	0.01	-0.02
VVVAGNRLSLIAR	R1	0.99	NA
AKVLEESLEIMSE	R2	0.00	0.01
HLVMELCEGGELF	A1	0.59	NA
DPKPDLTLGVPEH	B2	0.00	-0.02
RSVFDLYFRKNPF	B2	1.00	NA
WEVINLTEDNHPL	A1	0.00	0.80
PSIVDLDLNVFDK	B4	0.00	-0.02
HPLPSLSSSTKPH	A3	0.00	-0.02
ALEPDLTPFPANR	B4	0.00	NA
ACKVDLTREERSN	B3	0.00	0.76
KGSSTESIAVVEK	R1	0.00	0.01
LIESRTVVPLNTW	GID1a-SIM-A	0.62	-0.03
LLFPSSEPDLWRL	R3	0.00	-0.03
IKTMEDDYAVPVS	R3	0.00	0.06
TFLPDLTEIMTEI	B2	NA	NA
PFMIDLYPYYAYR	B1	0.68	0.04
NVLLSIESDKTDE	A2	0.00	0.06
IILPNLSSETTDL	A3	NA	NA
SFGNSEDPAPWSK	R3	0.00	-0.01
TRVLDLYNKNCSA	B2	0.28	0.03
QEPLLVEEEAACL	A2	NA	NA
VTPEESSLVVYLI	R3	0.23	NA
DRVIYLDSDIIVV	A2	NA	NA
DDVIVVESEERAR	A2	0.00	NA
EPIYSPSSDDHNL	A3	0.00	-0.03
DDVVTLSEDEEET	A2	NA	NA
RHLLVLSSDDNGC	A2	NA	NA
KRILKLNEGFDRF	A1	0.98	NA
LPPVDLTDLSSST	B1	NA	NA
KRVVDLCAAPGSW	B3	0.00	-0.04
DDDDDDDPDYVEE	R3	NA	NA
GFLLRISKQDIKK	A1	0.44	-0.01
ESKLDLTDTIKDG	B2	0.00	-0.05
AKMIDLSDNELTG	B3	0.00	-0.03
VAISSSSVKVLKI	R3	0.60	-0.03
RNMITMSEEVANS	A2	NA	NA
AGKKSDSLPVFDD	R3	0.00	-0.01

Peptide	Class	AtSUM1	HsSUM1
PEHIDLDQISVIH	B3	0.00	-0.02
IPCYDLTSSAPFL	B4	NA	0.94
DDVLDLSWSRSQL	B4	0.64	NA
MAPIKINGDKKLY	A1	0.00	-0.04
DNEKECIIVWEKK	M-IR2	0.00	0.23
LIVPRFEEDYWIR	A3	0.04	NA
LAVVVVSESYPIS	A2	NA	NA
VPKYDLSCLPSNI	B3	0.00	0.08
IDKYDLYGQVAMP	B1	0.00	0.00
PSEVDLYSSGENG	B1	NA	NA
SRPIPFDDDNKEV	A3	0.10	-0.02
RSLLEVEESLIQE	A3	NA	NA
SRGSSDDISVMLI	R2	NA	NA
PKFVDLTAPDHRP	B1	NA	NA
QKIVVVDSEMPSG	A2	0.00	-0.03
ALYVRFSSSCEKT	A3	0.00	0.03
KTTYDLSKKISWQ	B3	0.00	-0.03
TDSYDLFVISLIT	B3	0.22	NA
DVPLFLDEDVPLL	A3	NA	NA
VGEADLSVARRKN	B4	0.23	0.35
TDIISIEDSESED	A2	NA	NA
GTPVDLTRPSKKV	B1	0.01	0.04
RAMVDLDNAGVRG	B3	0.00	0.00
QEEEEEEVVVIGD	R2	NA	NA
LPLSRSRVLMVMN	R1	1.05	NA
IKAADLTSGKGKI	B2	0.09	0.02
VEIVDVKCGNPDR	A1	0.00	0.80
VAHLDLSKVFLQN	B4	0.09	0.00
NMKVDLTKLDMAA	B1	0.00	-0.03
MFPLLVCTRESTP	A1	NA	NA
QVMVAIDESECSK	A1	0.00	-0.03
WTQTDSEVEVMVI	R1	NA	NA
EQVYDLTEVKPNE	B3	0.00	-0.01
ISIICVCHGTSFS	A1	0.00	0.11
AVPIHTWVLISNF	ZmGID1a	0.65	-0.03
EVLLKLEEEYIEN	A3	0.00	-0.01
GRPMNVSDSSPLA	A2	NA	NA
FRLPPISSSADVA	A3	NA	NA
ASLLDICEKVLSL	A1	0.00	0.02
PFRKGNNILVMCD	R1	0.12	0.03
DPAVKGTLNVLNS	R1	0.00	0.19
ELIIDLTTSKIAS	B4	0.02	0.19
DVFWDLSSAKFGS	B4	0.59	-0.02
NDFLTLDESRLID	A3	0.00	0.04
EMKECTGLEVLPT	R1	0.00	0.00
SGFLDLYTDKDTT	B4	0.03	0.08
LEYVDLYLIHWPA	B1	0.00	-0.03
QELVTVRNEALHT	A1	0.24	-0.02
SGAMSSSLSLLVN	R2	0.79	NA
NLTPDLTTLDSFG	B2	NA	NA
EKTIDLTGRVGQE	B2	0.00	-0.03
RSESNGKLQLVYL	R1	0.05	-0.02
QFLVSVSESRSSI	A3	0.82	0.27
DSNFDEDLPLPPW	R3	NA	NA
YCLNKESVEVVKL	R1	0.04	NA
SSLLCLESEEARM	A3	NA	0.53
EALLDLTNSVVSS	B2	NA	NA
RALDSSDVELVKL	R2	0.00	0.03
DSILDLFLESVRI	B4	0.62	NA
EQLLALDSDKYSR	A2	0.00	-0.03
LVPKDESVKVLAE	R3	0.00	-0.04
PTIVEVSSSNMYD	A3	NA	NA
PSQFDSSISPMRM	R3	0.12	0.26

Peptide	Class	AtSUM1	HsSUM1
RLIIDLYGISKKP	B2	0.28	0.02
LIESRTVVPLNTW	GID1a-SIM-A	0.66	-0.04
WVLPEDDLILLAI	R2	0.15	NA
KFDKSESPRLVHR	R3	0.25	NA
LDPLVWDDEVAAY	A3	NA	NA
SMRSESDLELLNK	R2	0.00	-0.04
AQTFDLTPQNVDL	B2	NA	NA
GAGKTTQLRIITG	R1	0.75	NA
GEAMDLFQRMEKD	B4	0.08	0.00
RPILLLDSSAING	A3	NA	NA
RHGSSESISFPIR	R3	0.80	NA
INDLEEKMKLVKK	R1	0.01	-0.02
FRHSSSSLPILQQ	R2	0.67	0.33
KSVLGVSSEPSPR	A2	0.00	0.01
KMQKKQRMSLVVS	R1	0.05	0.00
SRKLEGKVALITG	R1	0.31	-0.01
NVLIDLYKKCGRI	B2	0.05	NA
AAIMAMDESVLDD	A2	NA	NA
KNMICFESSPNLM	A3	0.00	0.68
PSMVDLTTLKLLN	B3	0.80	0.15
QLGVDESYTLMVS	R3	NA	NA
PAILSVEDSKGNT	A2	0.00	-0.03
TSVIGIDSSLPAL	A2	NA	NA
ASILNPEESWKLC	A3	0.00	-0.03
AEILEVEEEGMSI	A2	NA	NA
EGELDLTSDPWPQ	B2	NA	NA
EIRERDGLFLMLQ	R1	0.87	NA
ACLSSSSVKVIKI	R2	0.02	-0.03
GCLVDLYARAGRL	B2	0.40	NA
RNELESDYVIVCF	R3	NA	NA
NNVVVMDSDSAVK	A2	0.00	0.07
TTMLSLSEDQHII	A2	0.00	-0.01
RVNIEDDLSVLGN	R2	NA	NA
SDVWDLTLRSHEL	B2	0.06	-0.01
YRYVDLTVINGNV	B1	NA	NA
PCLVPLGTSSIEN	A1	NA	0.30
SDKADSRIGLVIG	R1	0.28	0.44
LKAYDLTNLSEFL	B1	0.13	0.00
AVFIPIEEEEDCP	A3	NA	NA
STAVSDSVLLLFG	R2	0.74	NA
LKVLVVGGGDGGV	A1	0.05	-0.01
RVLLNICQEAFEG	A1	NA	NA
LNTIDLSSNNFEG	B4	NA	NA
YSVVVVNKENANG	A1	0.42	0.12
WALPEDELILLAI	R2	0.19	NA
ETTIDLSAKNNRG	B3	0.00	NA
NLLVDLYEDCCNG	B1	NA	NA
NRFYGLSSSSSSS	A3	0.76	0.48
FPLPDLTVTHENV	B2	0.00	-0.03
DWKYDLFPNPNFA	B3	0.56	-0.04
FGNHTEEVRLVWK	R1	0.00	0.01
PHFADLTSLRVID	B2	0.21	0.00
NIYVDEDMYILRD	R1	NA	NA
WKELDLESKIPNY	B4	0.00	-0.02
SSLIDLYCKCGFV	B1	0.00	-0.01
MIVMAPSDEAELF	A3	NA	NA
EAPNEEEVQLLLE	R2	NA	NA
VPIILVSSSSMAY	A2	0.87	NA
ISLSEDSVDVLED	R2	NA	NA
TSLITLRSNNIKE	A1	1.00	NA
STSIDLTRFMIQT	B1	0.98	NA
LGVLKVDESLRAE	A2	0.02	0.05
VRLWDLYTETPLF	B2	0.00	-0.02

Peptide	Class	AtSUM1	HsSUM1
FDLRSESFTFIGL	R3	0.75	NA
LVAKEDEIKMLKA	R2	0.00	-0.01
FGVYDLTTKTCTM	B1	0.00	0.00
EKFLDLTVSVICK	B2	0.00	0.01
YHKKQETLKILRQ	R1	0.49	NA
MYTIDLYDMTFRE	B1	0.05	0.00
FYVCDLEMLPCVL	B4	NA	NA
DVMVILDNHKTVP	A1	0.00	-0.01
AHQVDLYLLSQVA	B1	0.11	0.01
KSPLSLSDDEYVQ	A2	0.00	0.00
SSVIDSDMVLLLD	R1	NA	NA
SCLVTPSSEETND	A3	NA	0.32
KVEMDLYLVRKLV	B4	0.06	0.01
VGRADLTLSGSFA	B2	NA	NA
FWSHDEDPSPLPP	R3	0.00	0.00
KIIARKNLELVKW	R1	0.00	-0.02
LFYFDSSYRPVPL	R3	0.17	0.06
CPFFKMEEDHERI	A3	0.00	NA
TAMLKMEEEFVDS	A3	0.00	0.00
LESLQREVELLEQ	R1	0.00	-0.01
VKVWDLTAGKLLT	B4	0.00	0.01
LSRYDLFHGHLFL	B3	0.76	0.01
SIDGSDSLNYVNE	R3	NA	NA
DIPPDISDSDKYF	A3	0.00	-0.01
KMLIFLSESQSRQ	A3	0.39	0.10
SHMLDVEDDFEAF	A3	0.00	-0.01
VELLNMDDDDDGD	A2	NA	NA
LQVVDLYSNAISG	B1	NA	NA
TELPPLSESGLDW	A3	NA	NA
QTVFRPDDDVVIQ	A3	NA	NA
DDPFALESSTAGL	A3	NA	NA
GSKADLFLEPGDK	B4	0.01	-0.01
LSLLSLKTSLSGP	A1	0.58	0.09
FLSGHQTLQLMRK	R1	0.54	NA
LTLMYLKTESEVI	A1	0.05	0.01
GSLSSSDLPLLRP	R3	0.00	0.01
ATSDSSEPDLLWQ	R3	NA	NA
RLANDDDIMLLDW	R2	NA	NA
GRAIDLTRSECYE	B1	0.00	0.03
IPLVVVETRREVD	A1	0.04	0.00
IEFMDLYSYLIPV	B4	NA	NA
FILVCSSLDVLPH	R1	0.00	0.30
HHGHDEELHVLAV	R3	0.00	-0.01
KRKVDLTNDDVEG	B1	-0.01	0.00
FEESSESPPPMPN	R3	NA	NA
SLSVDLTTDAPNK	B1	0.01	0.05
WAEVEDSLIVINQ	R2	NA	NA
EYAEEEDMVLVDW	R3	NA	NA
CVRADLTDKCEAH	B2	0.00	-0.01
TPVIDLETVSGCD	B4	0.45	NA
LNIVCLCGHIFCW	A1	0.00	-0.02
IQRFDLTNINSAI	B2	0.68	0.11
IPLLQLSDSPRII	A2	0.51	0.03
DGFVDSDPIPIPI	R3	NA	NA
NSTIDLTKDPGHV	B2	0.01	-0.01
GTATSSELNLIDS	R2	0.00	-0.02
LHMFVWSSSASPV	A3	0.00	0.03
CGECDLFSMNEVL	S1ref753	0.39	0.28
GQKLNGSISVLGN	R1	0.00	0.07
DRLIDLEGREMQL	B4	NA	NA
IFWGEDELSMIRC	R1	0.00	-0.01
IPLLKERMNVLNP	R1	0.45	NA
SFIDSSSLWIVMP	R2	NA	NA

Peptide	Class	AtSUM1	HsSUM1
EVMIGYSDSGKDA	A3	0.00	-0.01
IISLDLSKSGLNG	B4	0.01	-0.01
PKFVDIDSDDKDP	A3	0.00	-0.01
KQGQGERINLLLE	R1	0.47	NA
KKTEENQLALMKP	R1	0.00	0.02
VAKASEEWGIFQV	R3	0.00	-0.01
MLSEDEEVRLLIS	R2	NA	NA
VVQKNGGIGLIYV	R1	0.10	0.00
GPYLTYEDSFLAI	A3	0.16	NA
LQHKNHDLELLCD	R1	0.60	0.06
ARRRGRGVSVLSG	R1	0.34	NA
RRRRSSSLALVPG	R2	0.37	NA
FAVYDLTTRTARH	B1	1.00	0.11
YTSQSSSMQILRA	R2	1.00	0.68
EERKKKTIVVMLP	R1	0.21	NA
IDMVLLDSSGTKI	A2	0.00	0.00
KYAGSEEMEVMLV	R2	0.00	0.05
KLAREESVILLPG	R2	0.03	-0.01
RDEIDLTQKKLPK	B1	0.00	0.03
EQVFDLTEVKPHE	B4	0.00	-0.01
LFVYDLETASAIY	B4	NA	NA
RVLPQSSVHIVVT	R1	0.84	0.45
KYMIDLEHRIKFL	B3	0.50	0.01
LPVIDLSLLHQPF	B3	0.80	0.03
LRSVDLSPASWLA	B3	0.00	-0.01
MMMMKMESEWVGA	A2	0.00	-0.03
SPEIDLTTNSEVT	B1	NA	NA
LVAPESSVDLVTV	R2	NA	NA
HRPPPPSSSTSRR	A3	0.39	NA
RTMIALEEKGVKY	A1	0.00	-0.03
CVPVIISDDIELP	A2	NA	0.37
NATVDLSFNNLTG	B3	0.48	NA
TQLADLYSSEGLS	B2	NA	NA
NRKVDLSMSKSKE	B4	0.03	0.02
ALFFDLTQAYRHT	B2	0.91	0.02
LQSLDLSQNRLSG	B4	0.84	0.05
QPKDDSSLRPYSQ	R3	0.00	0.16
DRLIAICEDLYAA	A1	NA	NA
MTLLKVDSEKVNK	A2	0.00	0.04
EAIKESRISIVIL	R1	0.39	0.07
NSIPSPSDETKDS	A3	-0.01	-0.02
LVLISDSVVLIST	R2	0.14	NA
VFYVDLSECVVCT	B3	0.53	-0.02
VAVVAMKEEDAGE	A1	0.00	-0.02
IMIVLLEESEDVQ	A2	NA	0.24
AFEKEDSVHIIDL	R2	0.00	-0.01
VSRCDSSVSVFSA	R3	NA	0.73
KKVVALESEIVEL	A2	-0.01	-0.01
GLLVHIEDSHLTR	A2	0.00	0.10
SDAVDLYEAMEKT	B1	0.00	-0.01
DSPLLLSSSHSLI	A2	0.13	0.00
FKTGENCMEMVNF	R1	0.00	0.03
MTFASDSMKLVVT	R3	0.00	-0.02
QGQPDLDIHRIRN	B4	0.32	NA
SKMMIVDDEYIII	A2	0.02	0.03
KLIFDIEEEPLQG	A3	0.10	0.04
LENKKRNVSVLRE	R1	0.01	NA
VSELDLYSPEEEI	B2	NA	NA
KKKLEEEVTILRS	R2	0.09	NA
AKAVDLTGKAPKS	B1	0.00	0.21
FDIYDLELTAVNN	B3	NA	NA
KSLVGMKKTLVFY	A1	0.03	-0.01
LNLVMLCQGVYDF	A1	0.00	-0.01

Peptide	Class	AtSUM1	HsSUM1
EVKCSEDIGLIGV	R2	-0.01	0.37
TALLDLYAKCGMI	B2	0.01	-0.01
RYSPDLTMFASGD	B2	NA	NA
SAVIPYEESGWYD	A3	NA	NA
WSMIDLYELIGRY	B1	0.05	-0.01
EKIVDIREEILRK	A1	0.30	NA
YNIWNVSSDDDDE	A3	NA	NA
AFADSSDVLLMIC	R2	NA	NA
QHLADLSFINRGG	B4	0.17	NA
ELPMREELGLMYV	R1	-0.01	-0.01
PELYEEDYNVLVD	R3	NA	NA
ERPISVEEEESGF	A2	NA	NA
KQIIDLYDQISKL	B1	0.10	0.02
RRECDLDEKDVFL	B4	0.01	0.98
MNFKSSSIVVMKA	R2	0.02	0.03
VSSPDLTKEVLLT	B2	0.01	-0.01
FIDEDEKVAVVFD	R1	0.88	-0.01
GASIDLEVCLPPE	B3	NA	NA
PAVLRLEDESRVL	A2	0.00	-0.01
FQGGEEDIIIIST	R3	0.15	NA
PVPLDEEVKLMIT	R3	0.00	0.01
WKKVDLTKIGIPS	B3	0.16	0.02
SLLLLISTSVTTS	A1	0.79	NA
INGTEEEVDLLRE	R2	-0.01	-0.01
FLQFDSDLTLIDP	R3	NA	NA
SGYLDLYQRLRNR	B2	0.91	NA
RWYLKPDDSSIFF	A3	0.05	0.01
VEADDDSVQMLDN	R2	NA	NA
NFKVSEEYGYPKN	R3	0.00	0.03
GSKVDLSDASMKG	B3	0.00	-0.01
REEWEESIKVYTE	R3	0.00	0.00
GPPLRVSSSQFPD	A2	NA	NA
VFIVHMSSSLAST	A2	0.18	0.01
TKDPSESMPVMVF	R2	0.00	-0.01
GIAVDLSDESAYA	B3	NA	NA
QHLSEDELQLLCE	R2	0.33	0.00
FHQIDLYGKLLGL	B4	0.85	0.00
EGVVVVDSEEIRR	A2	0.00	NA
LNVYDLTPVNNYL	B3	NA	NA
ESVPDRKVVILGA	R1	0.02	0.02
PDIFDLTVAKPSN	B2	0.00	0.00
DAILNIRKSYGFP	A1	0.96	NA
KETVDLENVPIEE	B3	0.00	-0.02
KNLVGVSDESYRL	A2	0.00	-0.01
APVSDDEVVPVED	R3	NA	NA
ANLVVMDDCELQG	A1	0.20	NA
GGLVVLTGTAVVA	A1	0.58	NA
QAMIDLDKTEKKS	A1	-0.01	0.00
GEKDESELLVVVG	R3	0.02	-0.01
IVAIDSEIYVLGG	R3	NA	NA
EICDSETVGIIMS	R1	NA	0.67
KVPVRLSESVKLW	A2	0.14	0.01
SPRVEESMSVVNN	R2	NA	NA
TFDSEESVDMVLH	R2	-0.01	-0.02
LLAVDLTDYCYRV	B1	0.00	-0.02
SLDASSDLFILNK	R2	0.10	0.03
LDELDLTDCLVLK	B2	0.06	0.02
YTDDEDDMMMVGD	R2	NA	NA
KALLGLEEDDLNG	A2	0.00	-0.01
YAEYDLYEIIRHH	B1	0.03	0.00
VDVYDLEDKMLFL	B3	0.00	0.00
AIYPDLTKNIEAF	B2	0.00	0.00
QAFGEDELAYLPD	R3	NA	NA

Peptide	Class	AtSUM1	HsSUM1
FAFVYMEDERDAE	A3	NA	NA
SVILDLTALYGIA	B4	0.13	NA
GDSFDLYYGLFLY	B2	NA	NA
GFYVQISDSLNST	A3	NA	NA
RTFVDLSTATMIV	B4	0.23	0.05
DKQADLYLDARPN	B2	-0.01	0.02
EWSLDSDYLLVGS	R3	NA	NA
ASPNEEELVVVGC	R2	NA	NA
TRADNSKMTLMHY	R1	0.00	-0.01
MVLMLPSDSARQW	A3	-0.01	-0.02
TCLISEELDFLKS	R3	0.00	0.12
LRLIGVEDSVGID	A2	NA	NA
PMMLQLGSGNEGN	A1	NA	NA
TSMIEFDSSESCE	A3	NA	NA
HPLMLVESESLTD	A2	NA	NA
SQPLPISSENKES	A2	0.00	-0.01
TTLIDLYVKCGCL	B1	0.01	-0.03
FGYTRKDVILIGV	R1	0.62	0.13
RPVSEDEVALMAK	R2	-0.01	0.01
GSNESSSMSIVMY	R2	NA	NA
DTVFDLTTAISKL	B4	0.06	0.10
FDLQSEEMNLLKE	R2	0.00	-0.01
GCSSSNTISLLLL	R1	0.72	NA
VMIMDVEKKGSII	A1	0.00	0.00
NQSVDLSSASDGN	B4	NA	NA
MNIASSSLPIPHN	R3	0.00	-0.01
SYLVTLSQQSGNV	A1	0.93	0.15
KCIIELTKDGDLR	A1	0.00	0.38
LPPSESEFIVFKL	R3	0.36	NA
MREHSSDLFMMTL	R3	0.00	0.00
ILTRDEELGVISD	R2	NA	NA
WAFIDLTAGPFSW	B1	NA	NA
ETQPEDSVHLVTW	R1	0.00	-0.01
SCLLESEVRILPD	R2	NA	0.91
TLVMDLTLCLSSI	B2	NA	NA
NDQNESSMLILQE	R1	NA	NA
RQIYDLYGEEGLK	B2	0.03	-0.01
GGIVDLEDIAGKA	B3	0.00	-0.01
LSLLLLSSSFSSV	A2	0.70	NA
ILTLDLFSQLAPV	B4	0.63	NA
IDEDESSLELIQI	R3	NA	NA
GKLLDLSDDPLWT	A3	0.00	-0.01
SEAVDLESVAVHE	B3	0.00	-0.01
EKALDLYYKMLNS	B2	0.47	0.00
QKKLDLTKDGAVS	B4	0.00	0.03
TGQDSSSVSIMNP	R2	NA	NA
KHWIDLTRILPLS	B1	0.93	0.12
KAEARDQMPLVIQ	R1	0.00	0.00
ALALDLFRKMEER	B4	0.88	NA
QDLMSLDDDILDF	A2	NA	NA
IAKQEEDIGFYAG	R3	0.08	-0.01
LSRSSDDWQYMGI	R3	NA	NA
IYMLPLGKGVSKA	A1	0.15	0.02
GKVGEDEFGYMLA	R3	-0.01	-0.01
IDVFDYSDEDDRV	A3	0.00	-0.01
AYILYLEESLRGL	A2	0.00	NA
SVLLRVKEDHDGA	A1	0.00	-0.01
DYHVDLESRLGKT	B4	0.00	0.00
RVLAKRGVRVVMA	R1	0.71	NA
RDVREEEITLMMA	R2	0.00	-0.01
DRLLAMSEDDLPY	A1	NA	NA
EKEKDDDVNIVIH	R2	0.01	0.00
EGFLDLTDVDIRL	B2	0.61	NA

Peptide	Class	AtSUM1	HsSUM1
ELLEKQQVFIVEG	R1	0.00	-0.01
FGYSSEDVQMVIE	R2	NA	NA
YLIMSIGESCDPV	A1	NA	NA
TGVFGPSSSSTNA	A3	NA	NA
CSSKSEGLNLLIY	R1	0.00	0.39
GRIVNLSSEAHRF	A2	0.31	NA
SWMMHVSSSLRLL	A2	1.03	NA
ETRSSSSMALITV	R2	0.42	0.07
QRSVDLYEDLIRC	B3	-0.01	-0.02
STPVALDDSSYSF	A2	NA	NA
LYMYDDEVPIPRK	R3	0.00	0.08
TTLVDLYAKCGDM	S1ref779	0.00	-0.01
VVVADLTKSKETT	B2	0.00	-0.02
LNVYDLTPMNVYG	B1	NA	NA
TPSIDLSVRKPSQ	B3	0.02	NA
AGIKEDDLLIMSD	R2	0.01	0.00
IPILPVEDDDIAM	A3	NA	NA
FAKPHKEVPMIFG	R1	-0.01	-0.01
VPILAVDSDGLVN	A3	0.71	NA
SLEIDLTRGKSVN	B1	0.01	NA
FSYFHNGVHVVSE	R1	0.03	0.01
HETGSTEIVVLCH	R1	0.00	0.00
LCLESDEVKMIGI	R3	-0.01	0.65
LAERTGHVMLLHL	R1	0.19	-0.01
RLVIIIGCSVLGF	A1	0.01	0.00
NWMKEDSLLFVHY	R3	0.39	-0.01
IKVSSSELSVLDE	R3	-0.01	-0.01
KHLVDLEMKLQIA	B3	0.00	0.00
VPAVSNGLIILYV	R1	0.52	NA
HPLLSLSSSPSSV	A2	NA	NA
GFPPFYSEEPLAT	A3	NA	NA
TYVIGLEDEEENK	A2	0.14	0.09
NKEADLEPGLDKA	B4	-0.01	-0.02
ELSGSSSLSVVFL	R2	0.72	NA
LLMEEDEIDFVAD	R3	NA	NA
QETKESDIKILRK	R3	0.04	0.05
KTSLSSSLMLVRL	R2	0.89	NA
LIGFSSSYSFVNF	R3	0.94	0.21
QQQVDLYDQHLAS	B4	-0.01	-0.01
PYVIKISRHHHRI	A1	0.60	NA
NKILTVNGEFPGP	A1	-0.01	-0.01
IDEVEEDMSLIGS	R1	NA	NA
DQVLHISTSPLHR	A1	0.36	0.04
DFLVEISDSNQTR	A2	1.02	NA
LKIIDLYVVFAVF	B3	0.65	0.09
YFRWDLYPYYRAF	B4	0.86	0.03
VTLADLTKKLKDW	B2	0.00	-0.01
EDEDEDDIPLVFK	R3	0.00	0.03
KYVLFLDDDVRLH	A3	0.04	0.01
EKVLEMEDSLESG	A2	-0.01	-0.01
AEVLGISSDSTII	A2	NA	NA
SSLILIKSEVAQS	A1	0.75	0.00
LSSQDSEPKPVNN	R3	-0.01	-0.01
LKIREEDLCVLVE	R2	-0.01	0.03
AAFVDLTPWHRFG	B1	0.43	NA
VWSEEDEISLLQA	R2	NA	NA
DGVISIESSSTSE	A2	NA	NA
MDEDEEEFELWLQ	R3	0.00	-0.01
STIVDLTKVGKYK	B1	0.00	-0.01
FGFYDLTTGEAHH	B1	0.00	0.00
RILPRPSDSVLKY	A3	0.40	0.05
HLFAESELRILID	R2	0.10	0.01
LVLESSDLMLMGF	R2	NA	NA

Peptide	Class	AtSUM1	HsSUM1
KIKESEDISILKA	R2	0.00	0.00
FYCLDLTLLLMGA	B2	NA	NA
IRRYDLTQDPLHS	B3	0.06	0.04
LETLDLSNNLALQ	B4	NA	NA
PLNPSENLLLLLQ	R1	0.76	NA
IPEVDLYKCEPWD	B1	-0.01	-0.02
RSPVLMSSSVFAL	A3	0.96	0.16
SSQVDLDNIDETE	B3	NA	NA
FAIWIMSSSQDDS	A3	0.07	-0.01
IVLIAVGNEITSF	A1	NA	NA
VEIYDLEENLVID	B3	NA	0.35
EKPLVIEDEQVIL	A3	-0.01	-0.01
VHIIDLDIMQGLQ	B3	0.06	0.09
QPQIDLTSNSGEY	B1	NA	NA
IGVVVVSDENGEP	A2	NA	NA
VTPVDLFYKRNHG	B3	1.06	0.01
KRLVMMGDDTWTQ	A1	0.04	-0.01
LPVIDLYRNIEHE	B3	0.00	-0.01
QSPIDLTDERVSL	B3	NA	NA
KRFVDEELVLVGT	R2	0.04	0.00
IPTIDLEEVNDQI	B3	0.00	0.36
AYGYSDEYSFVFK	R3	0.31	NA
KLPTRGSMKIIVK	R1	0.16	0.03
ISLYAISEERLPN	A3	0.02	0.04
VAPPSSSPMIVQK	R3	0.00	0.00
LVYIDLTSERLYK	B1	0.02	0.03
GQFVDLTRKLHTL	B1	0.96	NA
QRLMQLQQQQLLK	A1	0.14	NA
TEKYDLSSIRVVK	B3	0.99	NA
PSIADLDMDTYDK	B4	0.00	-0.01
HQVITVEENSAEH	A1	0.00	0.00
VQPIDLSGVGVPE	B3	NA	NA
KMIVLMSSDGQSF	A2	0.00	-0.01
RGRFSESLQLYKR	R3	0.98	NA
MRDMQDQLGILVR	R1	0.32	NA
SRSSESSVNILCL	R2	0.00	0.00
PLPVMYSSSLKRL	A3	1.00	0.10
IEKVDLSAKLTGQ	B4	0.00	-0.01
NVEDEDSIKIVET	R2	0.00	-0.01
LFILFVSTGRVIA	A1	0.74	0.34
QLPLDSSVVLVRD	R1	0.00	0.00
TSSCDLYRACGPF	B2	0.00	-0.02
PVIFSTNWLLVIN	lat62-1	NA	NA
KCLLKQVLLEDDM	lat12-1	0.00	0.24
GELLKIKSEKLPK	A1	0.01	0.05
PGLFDLSVSAEpYI	B4	NA	NA
PGLFDLpSVSAEYI	B4	NA	NA
PGLFDLpSVSAEpYI	B4	NA	NA
PGLFDLpSVpSAEYI	B4	0.14	NA
PGLFDLpSVpSAEpYI	B4	NA	NA
PGLFDLSVSAEYI	B4	0.37	NA
EKASGKKIPLVMA	R1	0.00	0.03
EKApSGKKIPLVMA	R1	0.00	0.09
FKCIDLDGDGVIT	B4	0.01	0.48
FKCIDLDGDGVIpT	B4	0.00	0.71
ISPIFISGGCEWF	A1	0.00	-0.06
IpSPIFISGGCEWF	A1	0.00	-0.05
ISPIFIpSGGCEWF	A1	0.00	-0.07
IpSPIFIpSGGCEWF	A1	-0.01	0.02
RALADLFLLpSNQR	B4	0.10	NA
RALADLFLLSNQR	B4	1.10	NA
LQKLDLSNNRLTG	B4	0.01	NA
LQKLDLpSNNRLTG	B4	0.00	NA

Peptide	Class	AtSUM1	HsSUM1
LQKLDLSNNRLpTG	B4	0.00	-0.02
LQKLDLpSNNRLpTG	B4	0.00	0.00
ASPLTIRSRLEEE	A1	0.03	0.00
ASPLpTIRSRLEEE	A1	0.00	0.00
ASPLTIRpSRLEEE	A1	0.00	-0.02
ApSPLpTIRSRLEEE	A1	0.04	0.00
ASPLpTIRpSRLEEE	A1	0.00	-0.01
ApSPLpTIRpSRLEEE	A1	0.00	-0.02
SNEPDLTPALLGP	B4	0.00	-0.02
pSNEPDLTPALLGP	B4	NA	NA
SNEPDLpTPALLGP	B4	NA	NA
pSNEPDLpTPALLGP	B4	NA	NA
DVFSENKIDLLPL	R1	0.00	-0.02
DVFpSENKIDLLPL	R1	0.00	-0.01
LGMIYVGSCNEEV	A1	NA	NA
LGMIpYVGSCNEEV	A1	NA	NA
LGMIYVGpSCNEEV	A1	1.05	0.22
LGMIpYVGpSCNEEV	A1	0.15	NA
ILLLFISSQVAIA	A1	0.59	0.03
ILLLFISpSQVAIA	A1	0.06	-0.02
ILLLFIpSpSQVAIA	A1	0.01	-0.03
ILLLFIpSSQVAIA	A1	0.01	-0.05
NLNSDKRLRLLSS	R1	0.38	NA
NLNSDKRLRLLSpS	R1	0.00	NA
NLNSDKRLRLLpSpS	R1	0.00	NA
NLNpSDKRLRLLSpS	R1	0.01	NA
NLNpSDKRLRLLpSpS	R1	0.00	NA
NLNpSDKRLRLLSS	R1	0.73	NA
SVEFDLDKTKRLL	B4	0.00	NA
pSVEFDLDKTKRLL	B4	0.00	NA
pSVEFDLDKpTKRLL	B4	0.01	NA
SVEFDLDKpTKRLL	B4	0.00	0.51
DLVLDLSAISEAG	B4	0.45	-0.01
DLVLDLpSAISEAG	B4	0.00	-0.03
DLVLDLpSAIpSEAG	B4	0.60	-0.02
DLVLDLSAIpSEAG	B4	0.03	-0.01
PQLLLIDTGSDLT	A1	0.00	-0.02
PQLLLIDTGSDLpT	A1	0.00	-0.02
PQLLLIDpTGSDLT	A1	0.00	0.00
PQLLLIDpTGSDLpT	A1	NA	NA
KPLLFLSSSSSFS	LAT17	0.81	0.11
KPLLFLSSSSSFS	LAT17	0.81	0.03
FEILDLTIGEWRT	LAT1	0.26	NA
PQLLLIDpTGpSDLT	A1	NA	NA
PQLLLIDpTGpSDLpT	A1	NA	NA
LFRFDLEAKCPPS	B4	0.01	-0.02
LFRFDLEAKCPPpS	B4	0.01	-0.02
LSPVIVQTSRWAN	A1	0.80	0.00
LSPVIVQTpSRWAN	A1	0.05	0.32
LSPVIVQpTSRWAN	A1	0.06	0.39
LSPVIVQpTpSRWAN	A1	0.00	NA
LpSPVIVQTpSRWAN	A1	0.00	NA
LpSPVIVQpTpSRWAN	A1	0.00	NA
LETMDLTEILRQK	B4	0.00	NA
LEpTMDLTEILRQK	B4	0.00	NA
LEpTMDLpTEILRQK	B4	0.00	NA
LETMDLpTEILRQK	B4	0.00	NA
GLKWDLSNTEMRF	B4	0.00	-0.05
GLKWDLpSNTEMRF	B4	0.00	NA
GLKWDLpSNpTEMRF	B4	0.00	NA
GLKWDLSNpTEMRF	B4	0.00	0.01
GLLLCVTKEDNIR	A1	0.00	0.46
GLLLCVpTKEDNIR	A1	0.00	0.17

Peptide	Class	AtSUM1	HsSUM1
ASYYDLSFHSISS	B4	0.09	0.09
ASYYDLSFHSIpSS	B4	0.70	0.15
ASYpYDLSFHpSISS	B4	0.00	0.00
ASYpYDLSFHSISpS	B4	0.00	-0.02
ApSYYDLpSFHSISpS	B4	0.20	0.04
ASYYDLSFHpSISS	B4	0.81	0.34
ESIVKIGEGTYGE	A1	0.00	-0.02
KPLLFLSSSSSFS	LAT17	0.85	0.13
FEILDLTIGEWRT	LAT1	0.33	0.02
KPLLFLSSSSSFS	LAT17	0.56	0.04
EpSIVKIGEGTYGE	A1	0.00	-0.01
EpSIVKIGEGpTYGE	A1	0.00	0.01
ESIVKIGEGpTpYGE	A1	0.00	-0.02
EpSIVKIGEGpTpYGE	A1	0.00	-0.02
ESIVKIGEGTpYGE	A1	0.00	-0.03
TKIEGQNLVVLGD	R1	0.00	-0.03
pTKIEGQNLVVLGD	R1	0.00	-0.05
KSPVWLKNDIERR	A1	0.00	NA
KpSPVWLKNDIERR	A1	0.03	NA
DRVLALQGNGSVN	A1	0.00	-0.08
DRVLALQGNGpSVN	A1	0.00	-0.11
AEEVDLSTDVQQW	B4	0.01	-0.10
AEEVDLpSTDVQQW	B4	0.00	-0.09
AEEVDLpSpTDVQQW	B4	0.00	-0.09
AEEVDLSpTDVQQW	B4	0.00	-0.06
GGHCDLELYPDFI	B4	0.00	-0.02
GGHCDLELpYPDFI	B4	0.00	0.00
YSVPDLDMLSVPD	B4	0.82	-0.03
YpSVPDLDMLSVPD	B4	0.00	-0.03
YSVPDLDMLpSVPD	B4	0.86	-0.02
pYpSVPDLDMLSVPD	B4	0.00	-0.03
pYpSVPDLDMLpSVPD	B4	-0.01	-0.03
YpSVPDLDMLpSVPD	B4	-0.01	-0.02
ITLLSLGSGEAPL	A1	0.04	-0.01
IpTLLSLGSGEAPL	A1	0.00	-0.04
ITLLpSLGSGEAPL	A1	0.01	-0.02
ITLLpSLGpSGEAPL	A1	NA	NA
IpTLLpSLGSGEAPL	A1	NA	NA
IpTLLpSLGpSGEAPL	A1	NA	NA
LIPVSVKEGDNVL	A1	0.00	-0.02
LIPVpSVKEGDNVL	A1	0.00	-0.02
MDMKEKKLTVIGT	R1	0.00	0.02
MDMKEKKLpTVIGT	R1	0.00	-0.02
MDMKEKKLTVIGpT	R1	0.00	-0.01
MDMKEKKLpTVIGpT	R1	0.00	-0.01
WSGGKTEVRLLFF	R1	0.35	0.12
WSGGKpTEVRLLFF	R1	0.72	0.09
WpSGGKpTEVRLLFF	R1	0.82	0.06
WpSGGKTEVRLLFF	R1	0.31	NA
LSLLRLGSTREPN	A1	0.71	NA
LSLLRLGpSTREPN	A1	0.01	NA
LpSLLRLGSTREPN	A1	0.88	NA
LpSLLRLGSpTREPN	A1	0.01	NA
LpSLLRLGpSpTREPN	A1	0.00	NA
LpSLLRLGpSTREPN	A1	0.03	NA
FELLDLSQPNFNN	B4	0.00	0.04
FELLDLpSQPNFNN	B4	-0.01	-0.02
PEPLRVGEKKEYD	A1	0.00	-0.03
PEPLRVGEKKEpYD	A1	0.00	-0.03
KKVVMVSEGFKHR	A1	0.00	NA
KKVVMVpSEGFKHR	A1	0.00	0.15
MTSFDLSILHIQI	B4	0.04	-0.01
MTSFDLpSILHIQI	B4	0.59	0.03

Peptide	Class	AtSUM1	HsSUM1
MpTSFDLSILHIQI	B4	0.18	-0.02
MTpSFDLpSILHIQI	B4	0.77	0.04
MpTpSFDLpSILHIQI	B4	0.87	0.13
MpTSFDLpSILHIQI	B4	0.67	0.14
TGLMGMNRGSLSF	A1	0.00	NA
TGLMGMNRGSLpSF	A1	0.00	-0.01
pTGLMGMNRGSLSF	A1	0.00	NA
TGLMGMNRGpSLpSF	A1	0.00	-0.01
pTGLMGMNRGSLpSF	A1	0.00	-0.01
pTGLMGMNRGpSLpSF	A1	NA	NA
pSMLFDLSRPARDL	B4	0.16	NA
SMLFDLpSRPARDL	B4	0.00	NA
pSMLFDLpSRPARDL	B4	0.00	NA
SMLFDLSRPARDL	B4	0.47	NA
NMPLLVHGEVTDP	A1	0.00	-0.05
NMPLLVHGEVpTDP	A1	0.00	-0.05
ENHLQHSLTIIPK	R1	0.00	NA
ENHLQHpSLTIIPK	R1	0.00	NA
ENHLQHSLpTIIPK	R1	0.00	NA
ENHLQHpSLpTIIPK	R1	0.00	NA
DTLVKMKNTVLHQ	A1	0.00	-0.04
DpTLVKMKNTVLHQ	A1	0.00	-0.01
DTLVKMKNpTVLHQ	A1	0.00	0.02
DpTLVKMKNpTVLHQ	A1	0.00	-0.02
SNTPTRSLSLISV	R1	0.00	0.03
pSNTPTRSLSLISV	R1	0.04	0.05
SNTPpTRSLSLISV	R1	0.05	0.11
SNTPTRSLpSLIpSV	R1	0.00	0.06
SNpTPpTRpSLSLISV	R1	NA	NA
SNTPpTRSLSLIpSV	R1	0.00	-0.01
SVLLDLRSGAKRA	A1	0.02	NA
pSVLLDLRSGAKRA	A1	0.13	NA
SVLLDLRpSGAKRA	A1	0.04	0.53
pSVLLDLRpSGAKRA	A1	0.01	NA
SEVIHLKEQLYEA	A1	0.00	-0.03
pSEVIHLKEQLYEA	A1	0.00	-0.03
SEVIHLKEQLpYEA	A1	0.00	-0.02
pSEVIHLKEQLpYEA	A1	0.00	-0.01
TGFLSRDVRLLSD	R1	0.07	0.11
TGFLSRDVRLLpSD	R1	0.00	-0.01
pTGFLSRDVRLLSD	R1	0.05	0.10
TGFLpSRDVRLLpSD	R1	0.00	0.03
pTGFLSRDVRLLpSD	R1	0.01	0.02
pTGFLpSRDVRLLpSD	R1	0.00	-0.01
GERLDLYEAARGK	B4	0.00	0.32
GERLDLpYEAARGK	B4	0.00	NA
KFILEKSVFLVVS	R1	0.74	0.05
KFILEKSVFLVVpS	R1	0.01	-0.03
KFILEKpSVFLVVpS	R1	0.04	-0.04
KFILEKpSVFLVVS	R1	0.63	0.01
HKVLHLNRTDTRL	A1	0.21	NA
HKVLHLNRTDpTRL	A1	0.07	NA
HKVLHLNRpTDpTRL	A1	0.00	NA
HKVLHLNRpTDTRL	A1	0.00	NA
LSIMAVCTSNERR	A1	0.00	0.87
LSIMAVCpTSNERR	A1	0.00	NA
LSIMAVCTpSNERR	A1	0.01	NA
LpSIMAVCpTSNERR	A1	0.01	NA
LpSIMAVCTpSNERR	A1	0.00	NA
LpSIMAVCpTpSNERR	A1	0.00	NA
CVLLLLSSHDNSR	A1	0.01	NA
CVLLLLSSHDNpSR	A1	0.00	NA
CVLLLLSpSHDNpSR	A1	0.00	NA

Peptide	Class	AtSUM1	HsSUM1
CVLLLLpSSHDNpSR	A1	0.01	NA
CVLLLLpSpSHDNpSR	A1	0.00	0.84
CVLLLLSpSHDNSR	A1	0.00	0.54
QSELDLSYGQRYQ	B4	0.18	-0.01
QSELDLSYGQRpYQ	B4	0.00	0.01
QpSELDLSpYGQRYQ	B4	0.01	NA
QSELDLpSYGQRpYQ	B4	0.00	0.01
QpSELDLpSpYGQRYQ	B4	0.01	NA
QpSELDLSYGQRYQ	B4	0.13	-0.01
AQLLEISESKVSQ	A1	0.00	-0.02
AQLLEIpSESKVSQ	A1	0.00	-0.03
AQLLEISEpSKVSQ	A1	0.00	-0.01
AQLLEIpSESKVpSQ	A1	0.00	-0.03
AQLLEIpSEpSKVSQ	A1	0.00	-0.02
AQLLEIpSEpSKVpSQ	A1	0.00	-0.03
RAIRESRIAVVVL	R1	0.12	0.09
RAIREpSRIAVVVL	R1	0.14	0.00
ELVMAIEEEFSIE	A1	0.00	-0.03
ELVMAIEEEFpSIE	A1	0.00	-0.01
LHELDNSVDIINQ	R1	0.00	0.03
LHELDNpSVDIINQ	R1	0.00	-0.01
AIVLDVGSGSVCH	A1	0.03	-0.01
AIVLDVGpSGSVCH	A1	0.57	0.01
AIVLDVGSGpSVCH	A1	0.99	0.01
AIVLDVGpSGpSVCH	A1	0.99	0.05
RKTFDLSYYQLVL	B4	0.05	0.01
RKpTFDLSYYQLVL	B4	0.03	-0.01
RKTFDLSYpYQLVL	B4	0.02	-0.05
RKTFDLpSYpYQLVL	B4	0.04	0.01
RKpTFDLSpYpYQLVL	B4	0.00	0.03
RKTFDLpSpYYQLVL	B4	0.26	0.10
LPQLDLFKSEIMS	B4	0.26	-0.03
LPQLDLFKpSEIMS	B4	0.00	-0.02
LPQLDLFKpSEIMpS	B4	0.00	-0.03
LPQLDLFKSEIMpS	B4	0.00	-0.02
EVGPETNVLVMGA	R1	0.00	-0.01
EVGPEpTNVLVMGA	R1	0.00	-0.03
FVVVVLSENYPTS	A1	0.00	0.00
FVVVVLSENYPTpS	A1	0.00	-0.02
FVVVVLSENpYPTS	A1	0.00	-0.03
FVVVVLpSENYPTpS	A1	0.01	-0.01
FVVVVLSENpYPTpS	A1	0.00	-0.02
FVVVVLpSENpYPpTS	A1	0.00	-0.02
PEMWDLFRREMIpS	B4	0.02	0.09
PEMWDLFRREMIS	B4	0.16	0.04
TILIDLDTKQVIE	A1	0.11	0.06
TILIDLDpTKQVIE	A1	0.00	-0.02
pTILIDLDpTKQVIE	A1	0.00	-0.02
pTILIDLDTKQVIE	A1	0.59	0.05
LRRLDLENVTKCL	B4	0.00	-0.02
LRRLDLENVpTKCL	B4	0.00	-0.03
IIKGTEKVLLIQE	R1	0.00	-0.03
IIKGpTEKVLLIQE	R1	0.00	-0.03
SPVESGRLAILAS	R1	0.03	-0.02
pSPVESGRLAILAS	R1	0.03	-0.02
SPVEpSGRLAILApS	R1	0.00	-0.02
pSPVEpSGRLAILAS	R1	0.01	-0.02
pSPVEpSGRLAILApS	R1	0.00	-0.01
SPVEpSGRLAILAS	R1	0.10	0.02
LCLVPMENTVGVA	A1	0.00	0.17
LCLVPMENpTVGVA	A1	NA	0.27
GKKMDLTYSVQWI	B4	0.00	-0.04
GKKMDLpTYSVQWI	B4	0.00	-0.03

Peptide	Class	AtSUM1	HsSUM1
GKKMDLTYpSVQWI	B4	0.00	-0.02
GKKMDLpTpYSVQWI	B4	0.00	-0.02
GKKMDLpTpYpSVQWI	B4	0.00	-0.02
GKKMDLTpYpSVQWI	B4	0.00	-0.03
GELLKIKpSEKLPK	A1	0.00	0.06
LAQLDLSYNKLTG	B4	0.00	-0.01
LAQLDLpSYNKLTG	B4	0.00	0.14
LAQLDLpSYNKLpTG	B4	0.00	-0.01
LAQLDLSpYNKLpTG	B4	0.00	-0.02
LAQLDLpSpYNKLpTG	B4	0.00	-0.02
LAQLDLSYNKLpTG	B4	0.00	-0.01
AKILGVDSRKSCV	A1	0.00	-0.02
AKILGVDSRKpSCV	A1	0.00	-0.02
AKILGVDpSRKpSCV	A1	0.00	-0.02
AKILGVDpSRKSCV	A1	0.00	-0.03
YLPPDLESEILSR	B4	0.00	0.01
YLPPDLESEILpSR	B4	0.00	NA
pYLPPDLESEILSR	B4	0.00	-0.02
YLPPDLEpSEILpSR	B4	0.00	NA
pYLPPDLEpSEILSR	B4	0.00	-0.01
pYLPPDLEpSEILpSR	B4	0.00	NA
DKVIDLSKDEKIE	B4	0.00	0.28
DKVIDLpSKDEKIE	B4	0.00	-0.03
AQLpYDLSGVPPER	B4	0.00	-0.02
AQLYDLpSGVPPER	B4	0.00	-0.02
AQLpYDLpSGVPPER	B4	0.00	-0.02
AQLYDLSGVPPER	B4	0.00	-0.01
IYIGTCEVGIVSV	R1	0.00	0.02
IYIGTCEVGIVpSV	R1	0.00	0.57
IYIGpTCEVGIVSV	R1	0.00	0.14
IpYIGTCEVGIVpSV	R1	0.00	0.62
IpYIGpTCEVGIVSV	R1	0.03	0.21
IpYIGpTCEVGIVpSV	R1	0.00	0.71
GCELDLSSAKGpYH	B4	0.00	0.31
GCELDLpSSAKGYH	B4	0.00	0.58
GCELDLpSpSAKGYH	B4	0.00	0.42
GCELDLSpSAKGpYH	B4	0.00	0.61
GCELDLpSpSAKGpYH	B4	0.00	0.57
GCELDLSSAKGYH	B4	0.00	0.30
LQTQESNIAIVGD	R1	0.00	-0.02
LQTQEpSNIAIVGD	R1	0.00	-0.02
LQpTQEpSNIAIVGD	R1	0.00	-0.02
LQpTQESNIAIVGD	R1	0.00	-0.02
FLAMTDEVRVIIV	R1	0.00	-0.01
FLAMpTDEVRVIIV	R1	0.00	0.02
PpTILDLFRNLGNV	B4	0.00	NA
PTILDLFRNLGNV	B4	0.03	NA
NGDDKTDIGMVVI	R1	0.00	-0.04
NGDDKpTDIGMVVI	R1	0.04	-0.04
GNIIGVDTGGVEK	A1	0.00	-0.04
GNIIGVDpTGGVEK	A1	0.00	-0.04
QQRLDLpYTSAAGL	B4	0.00	-0.04
QQRLDLYpTSAAGL	B4	0.00	-0.03
QQRLDLpYTpSAAGL	B4	0.00	-0.05
QQRLDLpYpTSAAGL	B4	0.00	-0.07
QQRLDLpYpTpSAAGL	B4	0.00	-0.04
QQRLDLYTSAAGL	B4	0.00	0.16
VETADLSDKALVpS	B4	0.00	-0.05
VETADLpSDKALVS	B4	0.00	-0.04
VEpTADLpSDKALVS	B4	0.00	-0.04
VETADLpSDKALVpS	B4	0.00	-0.03
VEpTADLpSDKALVpS	B4	0.00	-0.03
VETADLSDKALVS	B4	0.00	-0.03

Peptide	Class	AtSUM1	HsSUM1
ADVVVIGSGIGGL	A1	0.04	0.01
ADVVVIGpSGIGGL	A1	0.00	-0.02
LNLVYICQCNGLP	A1	0.02	-0.03
LNLVpYICQCNGLP	A1	0.00	-0.03
IVILEIRTEFGHK	A1	0.54	0.03
IVILEIRpTEFGHK	A1	0.07	0.00
ATLISGTVCLLVE	R1	0.00	-0.03
ATLIpSGTVCLLVE	R1	0.00	-0.02
ApTLISGpTVCLLVE	R1	0.00	-0.02
ApTLIpSGTVCLLVE	R1	0.00	-0.02
ApTLIpSGpTVCLLVE	R1	0.00	-0.02
ATLISGpTVCLLVE	R1	0.00	-0.03
ADVLDLCVDNpYES	B4	0.18	0.29
ADVLDLCVDNYEpS	B4	0.94	0.37
ADVLDLCVDNpYEpS	B4	0.03	0.03
ADVLDLCVDNYES	B4	1.07	0.11
VRILRVSNEGRES	A1	0.29	NA
VRILRVpSNEGRES	A1	0.00	-0.02
VRILRVSNEGREpS	A1	0.00	NA
VRILRVpSNEGREpS	A1	0.00	-0.03
DKRYDLFRTMSGK	B4	0.01	NA
DKRYDLFRpTMSGK	B4	0.00	NA
DKRYDLFRpTMpSGK	B4	0.00	NA
DKRpYDLFRTMpSGK	B4	0.00	NA
DKRpYDLFRpTMpSGK	B4	0.00	NA
DKRYDLFRTMpSGK	B4	0.00	NA
NpSKWDLTRQIANV	B4	0.00	NA
NSKWDLpTRQIANV	B4	0.00	NA
NpSKWDLpTRQIANV	B4	0.00	NA
NSKWDLTRQIANV	B4	0.03	0.06
KRGGGGRIILLTS	R1	0.00	0.01
KRGGGGRIILLpTS	R1	0.01	0.17
KRGGGGRIILLTpS	R1	0.09	0.14
KRGGGGRIILLpTpS	R1	0.17	NA
CVLLCVSQRKLQN	A1	0.01	0.05
CVLLCVpSQRKLQN	A1	0.00	0.02
RVLLHMCETSDLF	A1	0.01	0.18
RVLLHMCEpTSDLF	A1	0.00	0.23
RVLLHMCETpSDLF	A1	0.01	0.31
RVLLHMCEpTpSDLF	A1	0.00	0.14
CRSLDLTIISAED	B4	0.00	0.22
CRSLDLpTIISAED	B4	0.00	0.35
CRpSLDLpTIISAED	B4	NA	NA
CRSLDLpTIIpSAED	B4	0.00	0.75
CRpSLDLpTIIpSAED	B4	NA	0.30
CRpSLDLTIISAED	B4	0.19	NA
LHFTDHEISLLPR	R1	0.29	NA
LHFpTDHEISLLPR	R1	0.00	NA
LHFTDHEIpSLLPR	R1	0.04	NA
LHFpTDHEIpSLLPR	R1	0.00	NA
VDPWDLDTSIPLR	B4	0.00	NA
VDPWDLDpTSIPLR	B4	0.00	NA
VDPWDLDpTpSIPLR	B4	0.00	NA
VDPWDLDTpSIPLR	B4	0.00	NA
MPRQHGDLNIIYD	R1	0.03	-0.01
MPRQHGDLNIIpYD	R1	0.00	-0.03
AVDVTSELFLICL	R1	0.00	-0.03
AVDVpTSELFLICL	R1	0.00	-0.02
AVDVTpSELFLICL	R1	0.00	-0.02
AVDVpTpSELFLICL	R1	0.00	-0.03
GSLIVLRKDLGAP	A1	0.18	NA
GpSLIVLRKDLGAP	A1	0.28	0.20
KIEFDLDLLTLEP	B4	0.99	0.00

Peptide	Class	AtSUM1	HsSUM1
KIEFDLDLLpTLEP	B4	1.11	0.00
PSPLVVSGQYQDV	A1	0.06	-0.02
PSPLVVpSGQYQDV	A1	0.00	-0.02
PpSPLVVSGQYQDV	A1	0.00	-0.02
PSPLVVpSGQpYQDV	A1	0.00	-0.02
PpSPLVVpSGQpYQDV	A1	NA	NA
PpSPLVVSGQpYQDV	A1	NA	NA
KELLNMRSSSKRT	A1	0.03	0.02
KELLNMRSpSSKRT	A1	0.01	0.09
KELLNMRpSSSKRT	A1	0.16	0.07
KELLNMRpSSpSKRT	A1	0.02	0.12
KELLNMRpSpSSKRpT	A1	0.00	NA
KELLNMRSpSpSKRT	A1	0.01	0.04
VKTCHGSVSVVVY	R1	0.22	0.04
VKTCHGpSVSVVVY	R1	0.08	0.18
VKTCHGSVpSVVVY	R1	0.00	0.44
VKTCHGSVpSVVVpY	R1	0.00	0.63
VKTCHGpSVSVVVpY	R1	0.00	0.29
VKpTCHGSVpSVVVpY	R1	0.00	0.75
RYLNEDSLRMLLS	R1	0.02	NA
RYLNEDSLRMLLpS	R1	0.00	0.23
RpYLNEDpSLRMLLS	R1	0.00	NA
RpYLNEDSLRMLLpS	R1	0.00	0.01
RpYLNEDpSLRMLLpS	R1	0.00	NA
RpYLNEDSLRMLLS	R1	0.00	-0.01
VEPVRVKTELAEK	A1	0.00	-0.02
VEPVRVKpTELAEK	A1	0.00	-0.01
CSFFDLYLIYHSF	B4	0.00	-0.02
CSFFDLYLIYHpSF	B4	0.00	-0.01
CpSFFDLYLIYHpSF	B4	0.01	-0.02
CSFFDLpYLIpYHSF	B4	0.22	0.01
CSFFDLpYLIpYHpSF	B4	0.07	0.02
CSFFDLYLIpYHSF	B4	-0.01	0.00
ITSLDLSSSGLTG	B4	0.00	-0.02
ITpSLDLSSSGLTG	B4	0.00	-0.01
IpTSLDLSSSGLpTG	B4	0.00	-0.02
ITSLDLSpSSGLpTG	B4	0.00	-0.02
ITSLDLpSpSpSGLTG	B4	0.00	-0.02
IpTSLDLSSSGLTG	B4	0.04	-0.03
PYVTEDELKLMLR	R1	1.14	NA
PpYVTEDELKLMLR	R1	0.70	NA
PYVpTEDELKLMLR	R1	0.00	-0.01
PpYVpTEDELKLMLR	R1	0.00	-0.02
WAVVLLESEPEVL	A1	0.00	-0.03
WAVVLLEpSEPEVL	A1	0.00	-0.01
VGPVDLSSSAWSN	A1	0.00	-0.06
VGPVDLSSpSAWSN	A1	0.00	-0.05
VGPVDLSpSSAWpSN	A1	0.00	-0.08
VGPVDLpSpSSAWSN	A1	0.00	-0.08
VGPVDLpSpSSAWpSN	A1	0.00	-0.05
VGPVDLpSSSAWSN	A1	0.00	-0.03
ESMADLSLKpTNVP	B4	0.00	-0.02
ESMADLpSLKTNVP	B4	0.00	-0.02
ESMADLpSLKpTNVP	B4	0.00	-0.03
EpSMADLSLKpTNVP	B4	0.00	-0.02
EpSMADLpSLKpTNVP	B4	0.00	-0.02
ESMADLSLKTNVP	B4	0.00	-0.02
EVILCVDNRQNMY	A1	0.00	0.00
EVILCVDNRQNMpY	A1	0.00	0.08
NFKHSHQISVLVA	R1	0.40	-0.01
NFKHpSHQISVLVA	R1	0.84	0.01
NFKHSHQIpSVLVA	R1	0.31	-0.01
NFKHpSHQIpSVLVA	R1	0.08	-0.01

Peptide	Class	AtSUM1	HsSUM1
TDQIDLSKRSDST	B4	0.00	-0.02
TDQIDLSKRSDSpT	B4	0.00	-0.01
pTDQIDLSKRSDST	B4	0.00	-0.02
TDQIDLSKRpSDpST	B4	0.00	-0.01
TDQIDLpSKRSDpSpT	B4	0.00	-0.02
TDQIDLSKRSDpSpT	B4	0.00	-0.01
QGSLDLSAINPNQ	B4	0.06	-0.01
QGpSLDLSAINPNQ	B4	0.00	-0.02
QGpSLDLpSAINPNQ	B4	0.00	-0.03
QGSLDLpSAINPNQ	B4	0.00	-0.02
GSTADLTLDIASR	B4	0.01	-0.02
GSTADLpTLDIASR	B4	0.00	0.19
GSTADLTLDIApSR	B4	0.00	0.51
GSTADLpTLDIApSR	B4	0.00	0.33
GSpTADLpTLDIApSR	B4	0.44	NA
GpSTADLTLDIApSR	B4	0.00	0.04
VKLKRESLDLVNA	R1	0.00	0.13
VKLKREpSLDLVNA	R1	0.00	0.01
VGNQTREIKLLHR	R1	0.20	NA
VGNQpTREIKLLHR	R1	0.28	NA
LGQFKTCVLLLGN	R1	0.00	-0.02
LGQFKpTCVLLLGN	R1	0.00	-0.01
LKILDLpSFNKLNG	B4	0.03	NA
LKILDLSFNKLNG	B4	0.01	NA
DVNNKQQVTVVAE	R1	0.00	-0.01
DVNNKQQVpTVVAE	R1	0.00	-0.03
SHYYDLETLESSF	B4	0.00	-0.01
SHpYYDLETLESSF	B4	0.00	-0.01
SHYpYDLETLESSF	B4	0.00	-0.02
pSHYpYDLETLESSF	B4	0.00	-0.02
SHYpYDLEpTLESpSF	B4	0.01	-0.02
pSHYYDLETLESpSF	B4	0.03	-0.01
LPVLPVRRKTLLT	A1	0.32	NA
LPVLPVRRKpTLLT	A1	0.22	NA
LPVLPVRRKTLLpT	A1	0.15	NA
LPVLPVRRKpTLLpT	A1	0.00	NA
NLLVSSKLDVLKN	R1	0.00	0.04
NLLVSpSKLDVLKN	R1	0.00	0.00
NLLVpSpSKLDVLKN	R1	0.00	0.00
NLLVpSSKLDVLKN	R1	0.00	0.01
IPILDIDDSEFLH	A1	0.00	-0.01
IPILDIDDpSEFLH	A1	0.00	-0.01
VGLMDIGECDDAY	A1	0.00	0.05
VGLMDIGECDDApY	A1	0.00	0.04
LLPVEVKEQRVSN	A1	0.00	-0.04
LLPVEVKEQRVpSN	A1	0.00	-0.06
SRVLSIDTRVERA	A1	0.14	NA
pSRVLSIDTRVERA	A1	0.26	NA
SRVLpSIDTRVERA	A1	0.00	0.09
pSRVLSIDpTRVERA	A1	0.01	NA
pSRVLpSIDpTRVERA	A1	0.00	NA
pSRVLpSIDTRVERA	A1	0.00	0.00
EMEPTSSISLVAA	R1	0.00	-0.02
EMEPTpSSISLVAA	R1	0.00	-0.02
EMEPTSSIpSLVAA	R1	0.00	-0.02
EMEPpTpSSISLVAA	R1	0.00	-0.01
EMEPpTpSSIpSLVAA	R1	0.00	-0.02
EMEPTpSSIpSLVAA	R1	0.00	-0.02
ESSGKCGVAMMAS	R1	0.00	0.09
ESSGKCGVAMMApS	R1	0.00	0.22
KDYADLCFERFGD	B4	1.00	0.08
KDpYADLCFERFGD	B4	1.11	0.22
YELCDLFGMYMID	B4	0.05	-0.02

Peptide	Class	AtSUM1	HsSUM1
pYELCDLFGMYMID	B4	0.03	-0.01
YELCDLFGMpYMID	B4	0.01	0.20
pYELCDLFGMpYMID	B4	0.00	0.31
YEGAKTSIGIVPN	R1	0.00	0.08
YEGAKTpSIGIVPN	R1	0.00	-0.01
YEGAKpTpSIGIVPN	R1	0.00	-0.02
pYEGAKpTSIGIVPN	R1	0.00	-0.03
pYEGAKpTpSIGIVPN	R1	0.00	-0.02
YEGAKpTSIGIVPN	R1	0.00	0.01
YQAADLTKCGSQD	B4	0.00	-0.02
YQAADLpTKCGSQD	B4	0.00	-0.02
pYQAADLTKCGpSQD	B4	0.00	0.10
YQAADLpTKCGpSQD	B4	0.00	0.35
pYQAADLpTKCGpSQD	B4	0.00	0.42
YQAADLTKCGpSQD	B4	0.00	0.47
RSYSSSQLFIVII	R1	0.51	0.07
RSYpSSSQLFIVII	R1	0.45	0.04
RSpYSpSSQLFIVII	R1	0.34	0.02
RSpYSSpSQLFIVII	R1	0.40	0.02
RpSpYSSpSQLFIVII	R1	0.34	0.00
RpSYSSSQLFIVII	R1	0.69	0.00
pTSKPDLTSISSSP	B4	0.00	-0.02
TpSKPDLTSISSSP	B4	0.00	-0.02
TSKPDLTpSISSpSP	B4	0.00	-0.02
TSKPDLpTpSISSSP	B4	0.00	-0.01
pTSKPDLTpSISSpSP	B4	0.00	-0.02
TSKPDLTSISSSP	B4	0.00	-0.01
KRLREKTLEVIWQ	R1	0.08	-0.01
KRLREKpTLEVIWQ	R1	0.23	-0.03
VFGFDSSVHVVTG	R1	0.97	0.01
VFGFDpSSVHVVTG	R1	0.01	-0.02
VFGFDpSSVHVVpTG	R1	0.00	-0.02
VFGFDSpSVHVVpTG	R1	0.00	-0.03
VFGFDpSpSVHVVpTG	R1	0.00	-0.02
VFGFDSSVHVVpTG	R1	0.00	-0.03
QEIIGLTTKNANG	A1	0.00	-0.01
QEIIGLpTTKNANG	A1	0.00	0.00
QEIIGLTpTKNANG	A1	0.00	0.16
QEIIGLpTpTKNANG	A1	0.00	0.27
YALLTERIILVDN	R1	0.00	0.01
pYALLTERIILVDN	R1	0.00	-0.02
YALLpTERIILVDN	R1	0.00	-0.02
pYALLpTERIILVDN	R1	0.00	-0.04
SLATDHHLQMIGL	R1	0.00	-0.03
pSLATDHHLQMIGL	R1	0.00	0.03
SLApTDHHLQMIGL	R1	0.00	-0.02
pSLApTDHHLQMIGL	R1	0.00	-0.03
GQKVDLTRRIREV	B4	0.05	NA
GQKVDLpTRRIREV	B4	0.04	NA
MSVMEMSHRGKEF	A1	0.00	0.00
MpSVMEMSHRGKEF	A1	0.00	-0.01
MSVMEMpSHRGKEF	A1	0.00	0.02
MpSVMEMpSHRGKEF	A1	0.00	0.03
ESSLCKQLGIVPR	R1	0.00	0.53
ESpSLCKQLGIVPR	R1	0.00	0.63
EpSpSLCKQLGIVPR	R1	0.00	NA
EpSSLCKQLGIVPR	R1	0.00	0.70
VPRLDLDNRHFTE	B4	0.20	0.05
VPRLDLDNRHFpTE	B4	0.00	-0.01
VVSVSNGIPMLMR	R1	0.06	NA
VVpSVSNGIPMLMR	R1	0.00	NA
VVSVpSNGIPMLMR	R1	0.00	NA
VVpSVpSNGIPMLMR	R1	0.00	NA

Peptide	Class	AtSUM1	HsSUM1
NACADLpTTEELGK	B4	0.00	0.22
NACADLTpTEELGK	B4	0.00	0.46
NACADLpTpTEELGK	B4	0.00	0.28
NACADLTTEELGK	B4	0.00	0.36
GSPLLLEKRHNCL	A1	0.01	-0.01
GpSPLLLEKRHNCL	A1	0.04	0.04
WAYPDLDFIRWPI	B4	0.00	-0.03
WApYPDLDFIRWPI	B4	0.00	-0.03
RGMVSISGEPIQR	A1	0.00	NA
RGMVpSISGEPIQR	A1	0.00	0.05
RGMVSIpSGEPIQR	A1	0.00	0.17
RGMVpSIpSGEPIQR	A1	0.00	NA

B.3 Predicted SIM containing proteins

Table B.2: Top 500 predicted SIM containing proteins in *Arabidopsis*.

Accession ID	Short name	Description
AT5G39040	ABCB27, ALS1, AtALS1, ATTAP2, TAP2	transporter associated with antigen processing protein 2
AT4G39850	ABCD1, ACN2, AtABCD1, CTS, PED3, PXA1	peroxisomal ABC transporter 1
AT3G20320	ABCI15, TGD2	trigalactosyldiacylglycerol2
AT5G04895	ABO6	DEA(D/H)-box RNA helicase family protein
AT3G57330	ACA11	autoinhibited Ca2+-ATPase 11
AT4G37640	ACA2	calcium ATPase 2
AT3G05420	ACBP4, AtACBP4	acyl-CoA binding protein 4
AT5G49460	ACLB-2	ATP citrate lyase subunit B 2
AT4G33300	ADR1-L1	ADR1-like 1
AT1G12820	AFB3	auxin signaling F-box 3
AT2G24540	AFR	Galactose oxidase/kelch repeat superfamily protein
AT5G57090	AGR, AGR1, ATPIN2, EIR1, MM31, PIN2, WAV6	Auxin efflux carrier family protein
AT1G17260	AHA10	autoinhibited H(+)-ATPase isoform 10
AT5G62670	AHA11, HA11	H(+)-ATPase 11
AT4G30190	AHA2, HA2, PMA2	H(+)-ATPase 2
AT4G00730	AHDP, ANL2	Homeobox-leucine zipper family protein / lipid-binding START domain-containing protein
AT2G34680	AIR9	Outer arm dynein light chain 1 protein
AT1G69850	AIT1, ATNRT1:2, NRT1.2, NRT1:2, NTL1	nitrate transporter 1:2
AT5G35220	AMOS1, EGY1	Peptidase M50 family protein
AT4G32410	ANY1, AtCESA1, CESA1, RSW1	cellulose synthase 1
AT1G60780	AP1M2, HAP13	Clathrin adaptor complexes medium subunit family protein
AT1G56590	AP3M, ZIP4	Clathrin adaptor complexes medium subunit family protein
AT4G16130	ARA1, ATISA1, ISA1	arabinose kinase
AT2G44900	ARABIDILLO-1, ARABIDILLO1	ARABIDILLO-1
AT3G19180	ARC6H, ATCDP1, CDP1, PARC6	paralog of ARC6
AT1G68370	ARG1	Chaperone DnaJ-domain superfamily protein
AT2G16090	ARI2, ATARI2	RING/U-box superfamily protein
AT4G34940	ARO1	armadillo repeat only 1
AT3G27000	ARP2, ATARP2, WRM	actin related protein 2
AT1G18450	ARP4, ATARP4	actin-related protein 4
AT4G30510	ATATG18B, ATG18B	homolog of yeast autophagy 18 (ATG18) B
AT1G54710	ATATG18H, ATG18H	homolog of yeast autophagy 18 (ATG18) H
AT5G63810	AtBGAL10, BGAL10	beta-galactosidase 10
AT4G39400	ATBRI1, BIN1, BRI1, CBB2, DWF2	Leucine-rich receptor-like protein kinase family protein
AT2G46020	ATBRM, BRM, CHA2, CHR2	transcription regulatory protein SNF2, putative
AT2G02560	ATCAND1, CAND1, ETA2, HVE, TIP120	cullin-associated and neddylation dissociated
AT3G18480	AtCASP, CASP	CCAAT-displacement protein alternatively spliced product

Accession ID	Short name	Description
AT5G05170	ATCESA3, ATH-B, CESA3, CEV1, ELI1, IXR1, MRE1	Cellulose synthase family protein
AT5G17420	ATCESA7, CESA7, IRX3, MUR10	Cellulose synthase family protein
AT4G18780	ATCESA8, CESA8, IRX1, LEW2	cellulose synthase family protein
AT3G11130	AtCHC1, CHC1	Clathrin, heavy chain
AT3G06010	ATCHR12, CHR12	Homeotic gene regulator
AT2G13620	ATCHX15, CHX15, CHX15	cation/hydrogen exchanger 15
AT5G64660	ATCMPG2, CMPG2	CYS, MET, PRO, and GLY protein 2
AT5G53130	ATCNGC1, CNGC1	cyclic nucleotide gated channel 1
AT2G24610	ATCNGC14, CNGC14	cyclic nucleotide-gated channel 14
AT3G17700	ATCNGC20, CNBT1, CNGC20	cyclic nucleotide-binding transporter 1
AT5G54250	ATCNGC4, CNGC4, DND2, HLM1	cyclic nucleotide-gated cation channel 4
AT4G21670	ATCPL1, CPL1, FRY2, SHI4	C-terminal domain phosphatase-like 1
AT2G46700	ATCRK3, CRK3	CDPK-related kinase 3
AT5G17020	ATCRM1, ATXPO1, HIT2, XPO1, XPO1A	exportin 1A
AT4G31590	ATCSLC05, ATCSLC5, CSLC05, CSLC5	Cellulose-synthase-like C5
AT3G07330	ATCSLC06, ATCSLC6, CSLC06, CSLC6	Cellulose-synthase-like C6
AT4G07960	ATCSLC12, CSLC12, CSLC12	Cellulose-synthase-like C12
AT2G33100	ATCSLD1, CSLD1, CSLD1	cellulose synthase-like D1
AT3G03050	ATCSLD3, CSLD3, KJK, RHD7	cellulose synthase-like D3
AT4G38190	ATCSLD4, CSLD4	cellulose synthase like D4
AT1G02730	ATCSLD5, CSLD5, CSLD5, SOS6	cellulose synthase-like D5
AT4G02570	ATCUL1, AXR6, CUL1	cullin 1
AT1G26830	ATCUL3, ATCUL3A, CUL3, CUL3A	cullin 3
AT5G46210	ATCUL4, CUL4	cullin4
AT5G66750	ATDDM1, CHA1, CHR01, CHR1, DDM1, SOM1, SOM4	chromatin remodeling 1
AT1G55350	ATDEK1, DEK1, EMB1275, EMB80	calpain-type cysteine protease family
AT5G05980	ATDFB, DFB, FPGS1	DHFS-FPGS homolog B
AT1G73360	AtEDT1, ATHDG11, EDT1, HDG11	homeodomain GLABROUS 11
AT5G27640	ATEIF3B-1, ATTIF3B1, EIF3B, EIF3B-1, TIF3B1	translation initiation factor 3B1
AT3G56150	ATEIF3C-1, ATTIF3C1, EIF3C, EIF3C-1, TIF3C1	eukaryotic translation initiation factor 3C
AT1G79940	ATERDJ2A	DnaJ / Sec63 Brl domains-containing protein
AT1G66340	AtETR1, EIN1, ETR, ETR1	Signal transduction histidine kinase, hybrid-type, ethylene sensor
AT1G33390	ATFAS4, FAS4	RNA helicase family protein

Accession ID	Short name	Description
AT4G35930	AtFBS4, FBS4	F-box family protein
AT2G30390	ATFC-II, FC-II, FC2	ferrochelatase 2
AT5G15250	ATFTSH6, FTSH6	FTSH protease 6
AT3G01640	ATGLCAK, GLCAK	glucuronokinase G
AT1G42540	ATGLR3.3, GLR3.3	glutamate receptor 3.3
AT1G05200	ATGLR3.4, GLR3.4, GLUR3	glutamate receptor 3.4
AT4G01950	ATGPAT3, GPAT3	glycerol-3-phosphate acyltransferase 3
AT5G07830	AtGUS2, GUS2	glucuronidase 2
AT2G34710	ATHB-14, ATHB14, PHB, PHB-1D	Homeobox-leucine zipper family protein / lipid-binding START domain-containing protein
AT1G52150	ATHB-15, ATHB15, CNA, ICU4	Homeobox-leucine zipper family protein / lipid-binding START domain-containing protein
AT4G37270	ATHMA1, HMA1	heavy metal atpase 1
AT5G52640	ATHS83, AtHsp90-1, ATHSP90.1, HSP81-1, HSP81.1, HSP83, HSP90.1	heat shock protein 90.1
AT5G56000	AtHsp90.4, Hsp81.4	HEAT SHOCK PROTEIN 81.4
AT2G04030	AtHsp90.5, AtHsp90C, CR88, EMB1956, Hsp88.1, HSP90.5	Chaperone protein htpG family protein
AT5G52910	ATIM	timeless family protein
AT3G16630	ATKINESIN-13A, KINESIN-13A	P-loop containing nucleoside triphosphate hydrolases superfamily protein
AT5G05810	ATL43	RING/U-box superfamily protein
AT1G62830	ATLSD1, ATSWP1, LDL1, LSD1, SWP1	LSD1-like 1
AT3G44200	ATNEK6, IBO1, NEK6	NIMA (never in mitosis, gene A)-related 6
AT1G60800	AtNIK3, NIK3	NSP-interacting kinase 3
AT1G73590	ATPIN1, PIN1	Auxin efflux carrier family protein
AT5G57880	ATPRD2, MPS1, PRD2	multipolar spindle 1
AT5G53890	AtPSKR2, PSKR2	phytosylfokine-alpha receptor 2
AT1G60190	AtPUB19, PUB19	ARM repeat superfamily protein
AT3G08850	ATRAPTOR1B, RAPTOR1, RAPTOR1B	HEAT repeat ;WD domain, G-beta repeat protein protein
AT3G49750	AtRLP44, RLP44	receptor like protein 44
AT3G05530	ATS6A.2, RPT5A	regulatory particle triple-A ATPase 5A
AT1G22620	ATSAC1	Phosphoinositide phosphatase family protein
AT3G10380	ATSEC8, SEC8	subunit of exocyst complex 8
AT3G01680	AtSEOR1, SEOb, SEOR1	CONTAINS InterPro DOMAIN/s: Mediator complex subunit Med28 (InterPro:IPR021640);
AT4G04920	AtSFR6, IEN1, MED16, SFR6	sensitive to freezing 6
AT1G09020	ATSNF4, SNF4	homolog of yeast sucrose nonfermenting 4
AT5G20280	ATSPS1F, SPS1F, SPSA1	sucrose phosphate synthase 1F
AT1G04920	ATSPS3F, SPS3F	sucrose phosphate synthase 3F
AT3G43190	ATSUS4, SUS4	sucrose synthase 4
AT1G05500	ATSYTE, NTMC2T2.1, NTMC2TYPE2.1, SYT5, SYTE	Calcium-dependent lipid-binding (CaLB domain) family protein
AT1G06950	ATTIC110, TIC110	translocon at the inner envelope membrane of chloroplasts 110
AT3G20780	ATTOP6B, BIN3, HYP6, RHL3, TOP6B	topoisomerase 6 subunit B
AT1G78580	ATTPS1, TPS1, TPS1	trehalose-6-phosphate synthase
AT4G17770	ATTPS5, TPS5, TPS5	trehalose phosphatase/synthase 5

Accession ID	Short name	Description
AT1G06410	ATTPS7, ATTPSA, TPS7, TPS7	trehalose-phosphatase/synthase 7
AT2G03530	ATUPS2, UPS2	ureide permease 2
AT1G14360	ATUTR3, UTR3	UDP-galactose transporter 3
AT3G59360	ATUTR6, UTR6	UDP-galactose transporter 6
AT2G05170	ATVPS11, VPS11	vacuolar protein sorting 11
AT1G54560	ATXIE, XIE	Myosin family protein with Dil domain
AT1G03190	ATXPD, UVH6	RAD3-like DNA-binding helicase protein
AT3G13750	BGAL1, BGAL1	beta galactosidase 1
AT1G77410	BGAL16	beta-galactosidase 16
AT4G36360	BGAL3	beta-galactosidase 3
AT5G20710	BGAL7	beta-galactosidase 7
AT2G28470	BGAL8	beta-galactosidase 8
AT5G24630	BIN4, MID	double-stranded DNA binding
AT4G03080	BSL1	BRI1 suppressor 1 (BSU1)-like 1
AT2G27210	BSL3	BRI1 suppressor 1 (BSU1)-like 3
AT1G05940	CAT9	cationic amino acid transporter 9
AT3G19820	CBB1, DIM, DIM1, DWF1, EVE1	cell elongation protein / DWARF1 / DIMINUTO (DIM)
AT1G65320	CBSX6	Cystathionine beta-synthase (CBS) family protein
AT5G44030	CESA4, IRX5, NWS2	cellulose synthase A4
AT2G02090	CHA19, CHR19, ETL1	SNF2 domain-containing protein / helicase domain-containing protein
AT5G14170	CHC1	SWIB/MDM2 domain superfamily protein
AT3G47860	CHL	chloroplastic lipocalin
AT1G05490	chr31	chromatin remodeling 31
AT2G18760	CHR8	chromatin remodeling 8
AT3G52080	chx28	cation/hydrogen exchanger 28
AT1G05750	CLB19, PDE247	Tetratricopeptide repeat (TPR)-like superfamily protein
AT5G60600	CLB4, CSB3, GCPE, HDS, ISPG	4-hydroxy-3-methylbut-2-enyl diphosphate synthase
AT3G02130	CLI1, RPK2, TOAD2	receptor-like protein kinase 2
AT1G06220	CLO, GFA1, MEE5	Ribosomal protein S5/Elongation factor G/III/V family protein
AT4G24460	CLT2	CRT (chloroquine-resistance transporter)-like transporter 2
AT2G22125	CSI1, POM2	binding
AT1G63900	DAL1, SP1	E3 Ubiquitin ligase family protein
AT5G36950	DEG10, DegP10	DegP protease 10
AT5G40200	DEG9, DegP9	DegP protease 9
AT5G66680	DGL1	dolichyl-diphosphooligosaccharide-protein glycosyltransferase 48kDa subunit family protein
AT2G47420	DIM1A	Ribosomal RNA adenine dimethylase family protein
AT5G66360	DIM1B	Ribosomal RNA adenine dimethylase family protein
AT1G18260	EBS5, HRD3A	HCP-like superfamily protein
AT1G61140	EDA16	SNF2 domain-containing protein / helicase domain-containing protein / zinc finger protein-related
AT1G79350	EMB1135	RING/FYVE/PHD zinc finger superfamily protein
AT3G18110	EMB1270	Pentatricopeptide repeat (PPR) superfamily protein
AT5G49930	emb1441	zinc knuckle (CCHC-type) family protein
AT1G79490	EMB2217	Pentatricopeptide repeat (PPR) superfamily protein
AT3G48470	EMB2423	embryo defective 2423
AT4G11150	emb2448, TUF, TUFF, VHA-E1	vacuolar ATP synthase subunit E1
AT1G20200	EMB2719, HAP15	PAM domain (PCI/PINT associated module) protein
AT1G13980	EMB30, GN, VAN7	sec7 domain-containing protein
AT5G18700	EMB3013, RUK	Protein kinase family protein with ARM repeat domain
AT5G64580	EMB3144	AAA-type ATPase family protein
AT4G31820	ENP, MAB4, NPY1	Phototropic-responsive NPH3 family protein
AT4G02680	EOL1	ETO1-like 1
AT5G01400	ESP4	HEAT repeat-containing protein
AT4G26750	EXT-like	hydroxyproline-rich glycoprotein family protein

Accession ID	Short name	Description
AT3G14270	FAB1B	phosphatidylinositol-4-phosphate 5-kinase family protein
AT1G71010	FAB1C	FORMS APLOID AND BINUCLEATE CELLS 1C
AT1G22770	FB, GI	gigantea protein (GI)
AT3G10390	FLD	Flavin containing amine oxidoreductase family protein
AT2G30950	FTSH2, VAR2	FtsH extracellular protease family
AT1G13440	GAPC-2, GAPC2	glyceraldehyde-3-phosphate dehydrogenase C2
AT1G16300	GAPCP-2	glyceraldehyde-3-phosphate dehydrogenase of plastid 2
AT3G01040	GAUT13	galacturonosyltransferase 13
AT2G46180	GC4	golgin candidate 4
AT1G55325	GCT, MAB2	RNA polymerase II transcription mediators
AT5G58960	GIL1	Plant protein of unknown function (DUF641)
AT1G31070	GlcNAc1pUT1	N-acetylglucosamine-1-phosphate uridylyltransferase 1
AT5G24280	GMI1	gamma-irradiation and mitomycin c induced 1
AT2G13650	GONST1	golgi nucleotide sugar transporter 1
AT1G07290	GONST2	golgi nucleotide sugar transporter 2
AT1G32750	GTD1, HAC13, HAF01, HAF1, TAF1	HAC13 protein (HAC13)
AT1G64990	GTG1	GPCR-type G protein 1
AT1G10760	GWD, GWD1, SEX1, SOP, SOP1	Pyruvate phosphate dikinase, PEP/pyruvate binding domain
AT5G46880	HB-7, HDG5	homeobox-7
AT1G05230	HDG2	homeodomain GLABROUS 2
AT4G01690	HEMG1, PPO1, PPOX	Flavin containing amine oxidoreductase family
AT2G06990	HEN2	RNA helicase, ATP-dependent, SK12/DOB1 protein
AT1G63440	HMA5	heavy metal atpase 5
AT3G05040	HST, HST1	ARM repeat superfamily protein
AT1G64790	ILA	ILITYHIA
AT5G13460	IQD11	IQ-domain 11
AT3G49260	iqd21	IQ-domain 21
AT2G26180	IQD6	IQ-domain 6
AT4G38440	IYO	LOCATED IN: chloroplast;
AT3G13682	LDL2	LSD1-like2
AT1G02910	LPA1	tetratricopeptide repeat (TPR)-containing protein
AT1G14030	LSMT-L	Rubisco methyltransferase family protein
AT4G35920	MCA1	PLAC8 family protein
AT2G20980	MCM10	minichromosome maintenance 10
AT5G46280	MCM3	Minichromosome maintenance (MCM2/3/5) family protein
AT2G16440	MCM4	Minichromosome maintenance (MCM2/3/5) family protein
AT5G44635	MCM6	minichromosome maintenance (MCM2/3/5) family protein
AT4G02060	MCM7, PRL	Minichromosome maintenance (MCM2/3/5) family protein
AT2G14050	MCM9	minichromosome maintenance 9
AT2G14820	MEL3, NPY2	Phototropic-responsive NPH3 family protein
AT5G17520	MEX1, RCP1	root cap 1 (RCP1)
AT2G29990	NDA2	alternative NAD(P)H dehydrogenase 2
AT4G28220	NDB1	NAD(P)H dehydrogenase B1
AT1G30010	nMAT1	Intron maturase, type II family protein
AT2G43040	NPG1	tetratricopeptide repeat (TPR)-containing protein
AT4G28600	NPGR2	no pollen germination related 2
AT5G43050	NPQ6	Protein of unknown function (DUF565)
AT1G14850	NUP155	nucleoporin 155
AT3G57430	OTP84	Tetratricopeptide repeat (TPR)-like superfamily protein
AT2G48120	PAC	pale cress protein (PAC)
AT4G14210	PDE226, PDS, PDS3	phytoene desaturase 3
AT3G06960	PDE320, TGD4	pigment defective 320
AT3G25800	PDF1, PP2AA2, PR 65	protein phosphatase 2A subunit A2
AT4G04890	PDF2	protodermal factor 2
AT5G64070	PI-4KBETA1, PI4KBETA1	phosphatidylinositol 4-OH kinase beta1
AT1G72560	PSD	ARM repeat superfamily protein
AT5G42340	PUB15	Plant U-Box 15

Accession ID	Short name	Description
AT1G49780	PUB26	plant U-box 26
AT1G31830	PUT2	Amino acid permease family protein
AT3G19553	PUT5	Amino acid permease family protein
AT5G22750	RAD5, RAD5A	DNA/RNA helicase protein
AT3G55510	RBL	Noc2p family
AT3G51460	RHD4	Phosphoinositide phosphatase family protein
AT5G57280	RID2	S-adenosyl-L-methionine-dependent methyltransferases superfamily protein
AT5G44180	RLT2	Homeodomain-like transcriptional regulator
AT2G46710	ROPGAP3	Rho GTPase activating protein with PAK-box/P21-Rho-binding domain
AT4G29040	RPT2a	regulatory particle AAA-ATPase 2A
AT5G58290	RPT3	regulatory particle triple-A ATPase 3
AT1G58520	RXW8	lipases;hydrolases, acting on ester bonds
AT2G35800	SAMTL	mitochondrial substrate carrier family protein
AT3G56640	SEC15A	exocyst complex component sec15A
AT4G02350	SEC15B	exocyst complex component sec15B
AT1G21650	SECA2	Preprotein translocase SecA family protein
AT4G00800	SETH5	transducin family protein / WD-40 repeat family protein
AT5G48170	SLY2, SNE	F-box family protein
AT3G52490	SMXL3	Double Clp-N motif-containing P-loop nucleoside triphosphate hydrolases superfamily protein
AT1G03060	SPI	Beige/BEACH domain ;WD domain, G-beta repeat protein
AT5G42390	SPP	Insulinase (Peptidase family M16) family protein
AT2G02480	STI	AAA-type ATPase family protein
AT1G22150	SULTR1;3	sulfate transporter 1;3
AT2G19580	TET2	tetraspanin2
AT5G64510	TIN1	unknown protein;
AT3G24660	TMKL1	transmembrane kinase-like 1
AT3G01780	TPLATE	ARM repeat superfamily protein
AT4G31600	UTr7	UDP-N-acetylglucosamine (UAA) transporter family
AT3G13290	VCR	varicose-related
AT3G03660	WOX11	WUSCHEL related homeobox 11
AT2G35610	XEG113	xyloglucanase 113
AT4G11800		Calcineurin-like metallo-phosphoesterase superfamily protein
AT4G10080		unknown protein;
AT1G07590		Tetratricopeptide repeat (TPR)-like superfamily protein
AT5G43745		Protein of unknown function (DUF1012)
AT1G73950		Transmembrane Fragile-X-F-associated protein
AT3G07210		unknown protein;
AT3G14170		Plant protein of unknown function (DUF936)
AT2G28480		RNA-binding CRS1 / YhbY (CRM) domain protein
AT3G13670		Protein kinase family protein
AT3G18020		Pentatricopeptide repeat (PPR) superfamily protein
AT1G10330		Tetratricopeptide repeat (TPR)-like superfamily protein
AT1G12790		CONTAINS InterPro DOMAIN/s: RuvA domain 2-like (InterPro:IPR010994);
AT3G12020		P-loop containing nucleoside triphosphate hydrolases superfamily protein
AT2G26780		ARM repeat superfamily protein
AT1G45688		unknown protein;
AT1G62020		Coatomer, alpha subunit
AT2G32970		unknown protein;
AT5G06130		chaperone protein dnaJ-related
AT1G12380		unknown protein;
AT3G43240		ARID/BRIGHT DNA-binding domain-containing protein
AT3G49810		ARM repeat superfamily protein
AT4G24840		FUNCTIONS IN: molecular_function unknown;
AT5G58510		unknown protein;
AT1G30630		Coatomer epsilon subunit

Accession ID	Short name	Description
AT1G12500		Nucleotide-sugar transporter family protein
AT1G30290		Tetratricopeptide repeat (TPR)-like superfamily protein
AT5G42560		Abscisic acid-responsive (TB2/DP1, HVA22) family protein
AT4G02750		Tetratricopeptide repeat (TPR)-like superfamily protein
AT4G21660		proline-rich spliceosome-associated (PSP) family protein
AT4G22990		Major Facilitator Superfamily with SPX (SYG1/Pho81/XPR1) domain-containing protein
AT3G11320		Nucleotide-sugar transporter family protein
AT5G40250		RING/U-box superfamily protein
AT1G22860		Vacuolar sorting protein 39
AT3G08650		ZIP metal ion transporter family
AT5G47940		unknown protein;
AT2G45540		WD-40 repeat family protein / beige-related
AT2G45990		unknown protein;
AT5G15610		Proteasome component (PCI) domain protein
AT3G02710		ARM repeat superfamily protein
AT3G56570		SET domain-containing protein
AT1G59710		Protein of unknown function (DUF569)
AT1G55535		unknown protein;
AT2G24240		BTB/POZ domain with WD40/YVTN repeat-like protein
AT1G33420		RING/FYVE/PHD zinc finger superfamily protein
AT5G35970		P-loop containing nucleoside triphosphate hydrolases superfamily protein
AT1G08760		Plant protein of unknown function (DUF936)
AT1G77460		Armadillo/beta-catenin-like repeat ; C2 calcium/lipid-binding domain (CaLB) protein
AT4G13330		S-adenosyl-L-methionine-dependent methyltransferases superfamily protein
AT1G06590		unknown protein;
AT5G03250		Phototropic-responsive NPH3 family protein
AT4G02400		U3 ribonucleoprotein (Utp) family protein
AT4G27680		P-loop containing nucleoside triphosphate hydrolases superfamily protein
AT1G28690		Tetratricopeptide repeat (TPR)-like superfamily protein
AT4G29960		unknown protein;
AT3G25805		unknown protein;
AT3G26782		Tetratricopeptide repeat (TPR)-like superfamily protein
AT5G48130		Phototropic-responsive NPH3 family protein
AT5G48660		B-cell receptor-associated protein 31-like
AT1G05020		ENTH/ANTH/VHS superfamily protein
AT2G47010		unknown protein;
AT5G14260		Rubisco methyltransferase family protein
AT5G28850		Calcium-binding EF-hand family protein
AT5G43310		COP1-interacting protein-related
AT1G32460		unknown protein;
AT4G15840		BTB/POZ domain-containing protein
AT4G37920		unknown protein;
AT2G01460		P-loop containing nucleoside triphosphate hydrolases superfamily protein
AT2G28315		Nucleotide/sugar transporter family protein
AT4G11120		translation elongation factor Ts (EF-Ts), putative
AT5G39250		F-box family protein
AT3G07510		unknown protein;
AT5G15710		Galactose oxidase/kelch repeat superfamily protein
AT2G43320		S-adenosyl-L-methionine-dependent methyltransferases superfamily protein
AT5G15070		Phosphoglycerate mutase-like family protein
AT2G47390		Prolyl oligopeptidase family protein
AT3G19990		unknown protein;
AT2G22120		RING/FYVE/PHD zinc finger superfamily protein

Accession ID	Short name	Description
AT5G38520		alpha/beta-Hydrolases superfamily protein
AT2G26270		FUNCTIONS IN: molecular_function unknown;
AT3G03940		Protein kinase family protein
AT1G02040		C2H2-type zinc finger family protein
AT1G60560		SWIM zinc finger family protein
AT4G24160		alpha/beta-Hydrolases superfamily protein
AT5G12260		BEST Arabidopsis thaliana protein match is: glycosyltransferase family protein 2 (TAIR:AT5G60700.1);
AT1G15290		Tetratricopeptide repeat (TPR)-like superfamily protein
AT2G17930		Phosphatidylinositol 3- and 4-kinase family protein with FAT domain
AT2G42910		Phosphoribosyltransferase family protein
AT5G44860		unknown protein;
AT1G68710		ATPase E1-E2 type family protein / haloacid dehalogenase-like hydrolase family protein
AT1G73920		alpha/beta-Hydrolases superfamily protein
AT5G43020		Leucine-rich repeat protein kinase family protein
AT5G65750		2-oxoglutarate dehydrogenase, E1 component
AT4G24610		unknown protein;
AT5G16210		HEAT repeat-containing protein
AT5G35430		Tetratricopeptide repeat (TPR)-like superfamily protein
AT5G59740		UDP-N-acetylglucosamine (UAA) transporter family
AT1G48360		zinc ion binding;nucleic acid binding;hydrolases, acting on acid anhydrides, in phosphorus-containing anhydrides
AT2G01690		ARM repeat superfamily protein
AT5G39450		F-box family protein
AT5G49960		unknown protein;
AT3G49350		Ypt/Rab-GAP domain of gyp1p superfamily protein
AT4G19006		Proteasome component (PCI) domain protein
AT3G60860		SEC7-like guanine nucleotide exchange family protein
AT4G31480		Coatomer, beta subunit
AT5G13500		unknown protein;
AT5G58160		actin binding
AT1G66330		senescence-associated family protein
AT3G03790		ankyrin repeat family protein / regulator of chromosome condensation (RCC1) family protein
AT5G22780		Adaptor protein complex AP-2, alpha subunit
AT1G14330		Galactose oxidase/kelch repeat superfamily protein
AT1G20540		Transducin/WD40 repeat-like superfamily protein
AT5G42740		Sugar isomerase (SIS) family protein
AT1G03440		Leucine-rich repeat (LRR) family protein
AT1G45150		unknown protein;
AT1G13820		alpha/beta-Hydrolases superfamily protein
AT2G42700		FUNCTIONS IN: molecular_function unknown;
AT1G27660		basic helix-loop-helix (bHLH) DNA-binding superfamily protein
AT2G25760		Protein kinase family protein
AT2G40400		Protein of unknown function (DUF399 and DUF3411)
AT2G16760		Calcium-dependent phosphotriesterase superfamily protein
AT2G29670		Tetratricopeptide repeat (TPR)-like superfamily protein
AT3G22800		Leucine-rich repeat (LRR) family protein
AT3G49142		Tetratricopeptide repeat (TPR)-like superfamily protein
AT4G01210		glycosyl transferase family 1 protein
AT5G05570		transducin family protein / WD-40 repeat family protein
AT1G73430		sec34-like family protein
AT3G12650		unknown protein;
AT3G24480		Leucine-rich repeat (LRR) family protein
AT5G19640		Major facilitator superfamily protein
AT1G17500		ATPase E1-E2 type family protein / haloacid dehalogenase-like hydrolase family protein

Accession ID	Short name	Description
AT2G02170		Remorin family protein
AT3G45850		P-loop containing nucleoside triphosphate hydrolases superfamily protein
AT4G34310		alpha/beta-Hydrolases superfamily protein
AT5G28350		Quinoprotein amine dehydrogenase, beta chain-like; RIC1-like guanyl-nucleotide exchange factor
AT1G19835		Plant protein of unknown function (DUF869)
AT1G70280		NHL domain-containing protein
AT5G42760		Leucine carboxyl methyltransferase
AT3G52870		IQ calmodulin-binding motif family protein
AT1G14390		Leucine-rich repeat protein kinase family protein
AT4G18820		AAA-type ATPase family protein
AT1G22850		SNARE associated Golgi protein family
AT1G50020		unknown protein;
AT2G15860		unknown protein;
AT3G44330		INVOLVED IN: protein processing;
AT5G06120		ARM repeat superfamily protein
AT4G34450		coatomer gamma-2 subunit, putative / gamma-2 coat protein, putative / gamma-2 COP, putative
AT1G07970		CONTAINS InterPro DOMAIN/s: Cytochrome B561-related, N-terminal (InterPro:IPR019176);
AT2G40280		S-adenosyl-L-methionine-dependent methyltransferases superfamily protein
AT4G39050		Kinesin motor family protein
AT5G08420		RNA-binding KH domain-containing protein
AT5G61450		P-loop containing nucleoside triphosphate hydrolases superfamily protein
AT1G03010		Phototropic-responsive NPH3 family protein
AT2G26690		Major facilitator superfamily protein
AT2G41770		Protein of unknown function (DUF288)
AT4G16470		Tetratricopeptide repeat (TPR)-like superfamily protein
AT5G11720		Glycosyl hydrolases family 31 protein
AT1G63850		BTB/POZ domain-containing protein
AT3G10210		SEC14 cytosolic factor family protein / phosphoglyceride transfer family protein
AT1G50450		Saccharopine dehydrogenase
AT3G55060		unknown protein;
AT4G21300		Tetratricopeptide repeat (TPR)-like superfamily protein
AT3G56120		S-adenosyl-L-methionine-dependent methyltransferases superfamily protein
AT4G25030		unknown protein;
AT5G16610		unknown protein;
AT1G50440		RING/FYVE/PHD zinc finger superfamily protein
AT2G40980		Protein kinase superfamily protein
AT1G70160		unknown protein;
AT5G08720		CONTAINS InterPro DOMAIN/s: Streptomyces cyclase/dehydrase (InterPro:IPR005031);
AT2G03270		DNA-binding protein, putative
AT2G27610		Tetratricopeptide repeat (TPR)-like superfamily protein
AT1G71060		Tetratricopeptide repeat (TPR)-like superfamily protein
AT1G67930		Golgi transport complex protein-related
AT1G06890		nodulin MtN21 /EamA-like transporter family protein
AT3G01720		unknown protein;
AT4G32750		unknown protein;
AT4G04670		Met-10+ like family protein / kelch repeat-containing protein
AT5G49665		Zinc finger (C3HC4-type RING finger) family protein
AT2G32415		Polynucleotidyl transferase, ribonuclease H fold protein with HRDC domain
AT1G76140		Prolyl oligopeptidase family protein
AT5G19540		unknown protein;

Accession ID	Short name	Description
AT3G48770		DNA binding;ATP binding
AT5G51150		Mitochondrial import inner membrane translocase subunit Tim17/Tim22/Tim23 family protein
AT1G01930		zinc finger protein-related
AT2G25430		epsin N-terminal homology (ENTH) domain-containing protein / clathrin assembly protein-related
AT4G38200		SEC7-like guanine nucleotide exchange family protein
AT4G36180		Leucine-rich receptor-like protein kinase family protein
AT5G37490		ARM repeat superfamily protein
AT1G53345		unknown protein;
AT2G04360		unknown protein;
AT3G59340		Eukaryotic protein of unknown function (DUF914)
AT5G16680		RING/FYVE/PHD zinc finger superfamily protein
AT4G30400		RING/U-box superfamily protein
AT4G32272		Nucleotide/sugar transporter family protein
AT3G05990		Leucine-rich repeat (LRR) family protein
AT4G15890		binding
AT4G19380		Long-chain fatty alcohol dehydrogenase family protein
AT3G17740		unknown protein;
AT3G54190		Transducin/WD40 repeat-like superfamily protein
AT5G08540		unknown protein;
AT5G63100		S-adenosyl-L-methionine-dependent methyltransferases superfamily protein
AT1G12470		zinc ion binding
AT4G02900		ERD (early-responsive to dehydration stress) family protein
AT3G01580		Tetratricopeptide repeat (TPR)-like superfamily protein
AT3G04480		endoribonucleases
AT2G35840		Sucrose-6F-phosphate phosphohydrolase family protein
AT3G17030		Nucleic acid-binding proteins superfamily
AT4G37030		unknown protein;
AT2G22400		S-adenosyl-L-methionine-dependent methyltransferases superfamily protein
AT3G27550		RNA-binding CRS1 / YhbY (CRM) domain protein
AT3G28040		Leucine-rich receptor-like protein kinase family protein
AT4G28890		RING/U-box superfamily protein
AT2G33680		Tetratricopeptide repeat (TPR)-like superfamily protein
AT3G62360		Carbohydrate-binding-like fold
AT4G33970		Leucine-rich repeat (LRR) family protein
AT5G25265		unknown protein;
AT1G16860		Ubiquitin-specific protease family C19-related protein
AT4G13970		zinc ion binding
AT5G64270		splicing factor, putative
AT5G56220		P-loop containing nucleoside triphosphate hydrolases superfamily protein
AT5G43530		Helicase protein with RING/U-box domain
AT5G64090		FUNCTIONS IN: molecular_function unknown;
AT3G59910		Ankyrin repeat family protein
AT3G47530		Pentatricopeptide repeat (PPR) superfamily protein
AT2G38000		chaperone protein dnaJ-related
AT3G49650		P-loop containing nucleoside triphosphate hydrolases superfamily protein
AT3G50780		BEST Arabidopsis thaliana protein match is: BTB/POZ domain-containing protein (TAIR:AT1G63850.1);
AT3G07950		rhomboid protein-related
AT4G34220		Leucine-rich repeat protein kinase family protein
AT1G12800		Nucleic acid-binding, OB-fold-like protein
AT3G58480		calmodulin-binding family protein
AT2G32730		26S proteasome regulatory complex, non-ATPase subcomplex, Rpn2/Psmd1 subunit
AT5G02550		unknown protein;

Accession ID	Short name	Description
AT5G21070		unknown protein;
AT5G60700		glycosyltransferase family protein 2
AT2G01600		ENTH/ANTH/VHS superfamily protein
AT5G42930		alpha/beta-Hydrolases superfamily protein
AT3G13600		calmodulin-binding family protein
AT4G15820		BEST Arabidopsis thaliana protein match is: embryo defective 1703 (TAIR:AT3G61780.1);
AT2G34250		SecY protein transport family protein
AT5G11710		ENTH/VHS family protein
AT1G16220		Protein phosphatase 2C family protein
AT2G07360		SH3 domain-containing protein
AT5G65290		LMBR1-like membrane protein
AT5G11700		LOCATED IN: vacuole;
AT1G12600		UDP-N-acetylglucosamine (UAA) transporter family
AT2G39910		ARM repeat superfamily protein
AT4G32140		EamA-like transporter family
AT4G28080		Tetratricopeptide repeat (TPR)-like superfamily protein
AT5G66960		Prolyl oligopeptidase family protein

B.4 SIM containing protein gene ontology analysis

Table B.3: Full molecular function gene ontology analysis of SIM containing proteins.

Molecular function GO term	**Observed frequency (%)**	**Expected frequency (%)**	**Ratio**	***p*-value**
hydrolase activity, acting on acid anhydrides	12.05	6.54	1.84	7.89E-13
hydrolase activity, acting on acid anhydrides, in phosphorus-containing anhydrides	12.05	6.56	1.84	7.89E-13
nucleoside-triphosphatase activity	14.55	8.84	1.65	7.89E-13
pyrophosphatase activity	2.50	0.65	3.86	1.53E-12
cellulose synthase activity	4.09	0.57	7.24	2.59E-08
glucosyltransferase activity	1.36	0.09	14.99	2.59E-08
ATPase activity	11.59	5.68	2.04	9.94E-08
transporter activity	39.55	32.51	1.22	2.88E-07
substrate-specific transporter activity	5.00	2.23	2.24	6.97E-07
ATPase activity, coupled	12.05	6.07	1.98	4.69E-06
substrate-specific transmembrane transporter activity	1.14	0.11	10.26	6.78E-06
calmodulin binding	2.50	0.15	17.09	7.04E-06
hydrolase activity	2.95	0.64	4.58	1.10E-05
ion transmembrane transporter activity	12.50	6.96	1.80	2.60E-05
ATP binding	11.36	3.01	3.78	4.26E-05
adenyl ribonucleotide binding	11.36	3.03	3.75	4.49E-05
ATPase activity, coupled to transmembrane movement of ions, phosphorylative mechanism	11.14	2.87	3.87	4.62E-05
purine nucleoside binding	3.64	0.70	5.23	4.62E-05
adenyl nucleotide binding	11.59	5.66	2.05	4.62E-05
transferase activity, transferring hexosyl groups	1.14	0.14	7.98	4.62E-05
transmembrane transporter activity	1.59	0.34	4.68	4.62E-05
ATPase activity, coupled to transmembrane movement of ions	4.55	1.28	3.55	4.68E-05
binding	12.05	6.06	1.99	4.68E-05
nucleoside binding	1.36	0.13	10.45	4.68E-05
beta-galactosidase activity	11.14	2.99	3.72	4.90E-05
galactosidase activity	1.36	0.08	18.15	1.19E-04
helicase activity	9.55	3.55	2.69	1.59E-04
purine ribonucleoside triphosphate binding	2.50	0.64	3.90	2.64E-04
ribonucleotide binding	20.00	11.62	1.72	2.71E-04
purine ribonucleotide binding	2.50	0.64	3.90	2.71E-04
purine nucleotide binding	2.95	0.85	3.46	3.23E-04
DNA-dependent ATPase activity	12.05	4.97	2.42	5.76E-04
nucleotide-sugar transmembrane transporter activity	8.18	3.05	2.68	8.17E-04
nucleotide binding	0.91	0.04	22.99	8.17E-04
transferase activity, transferring glycosyl groups	4.09	1.65	2.48	9.17E-04
protein binding	2.95	0.85	3.48	1.03E-03
P-P-bond-hydrolysis-driven transmembrane transporter activity	5.00	1.94	2.58	1.68E-03
primary active transmembrane transporter activity	10.91	6.10	1.79	1.71E-03
ATPase activity, coupled to movement of substances	1.82	0.17	10.69	1.92E-03
ATPase activity, coupled to transmembrane movement of substances	8.86	3.84	2.31	1.92E-03
clathrin binding	6.59	1.70	3.89	2.05E-03
hydrolase activity, acting on acid anhydrides, catalyzing transmembrane movement of substances	5.00	1.26	3.96	2.05E-03

Molecular function GO term	**Observed frequency (%)**	**Expected frequency (%)**	**Ratio**	***p*-value**
active transmembrane transporter activity	6.36	2.18	2.92	4.72E-03
cation transmembrane transporter activity	47.50	36.72	1.29	4.93E-03
cation-transporting ATPase activity	2.05	0.24	8.48	5.40E-03
ion channel activity	12.05	6.56	1.84	8.95E-03
catalytic activity	12.05	6.09	1.98	9.83E-03

Table B.4: Full biological process gene ontology analysis of SIM containing proteins.

Biological process GO term	**Observed frequency (%)**	**Expected frequency (%)**	**Ratio**	***p*-value**
cellular component organization or biogenesis	13.93	5.13	2.71	3.12E-09
developmental process	17.12	7.22	2.37	3.12E-09
plant-type cell wall biogenesis	2.74	0.20	13.68	1.10E-07
cell wall biogenesis	2.97	0.25	11.76	1.10E-07
cellular component organization or biogenesis at cellular level	10.73	3.79	2.83	1.10E-07
cellular component organization	10.96	4.08	2.69	2.30E-07
developmental process involved in reproduction	9.36	3.23	2.90	4.38E-07
anatomical structure morphogenesis	5.94	1.49	3.99	8.93E-07
cellular developmental process	5.94	1.51	3.92	1.14E-06
cellular cell wall organization or biogenesis	3.65	0.60	6.04	3.36E-06
cell morphogenesis	4.57	0.99	4.61	3.69E-06
cellular component morphogenesis	4.57	0.99	4.61	3.69E-06
reproductive process	9.36	3.61	2.59	4.70E-06
cell morphogenesis involved in differentiation	3.42	0.56	6.06	6.11E-06
transport	14.16	6.94	2.04	8.01E-06
establishment of localization	14.16	6.99	2.02	9.67E-06
DNA-dependent DNA replication initiation	1.37	0.04	31.08	1.47E-05
cellular component organization at cellular level	7.53	2.77	2.72	2.30E-05
plant-type cell wall organization or biogenesis	2.97	0.47	6.28	2.30E-05
carbohydrate biosynthetic process	4.11	0.96	4.29	2.80E-05
cell wall organization or biogenesis	3.88	0.95	4.11	9.66E-05
organelle organization	6.16	2.16	2.85	9.98E-05
polysaccharide biosynthetic process	2.51	0.39	6.40	1.21E-04
disaccharide metabolic process	1.83	0.19	9.70	1.71E-04
embryo development	5.02	1.65	3.04	2.77E-04
polysaccharide metabolic process	2.97	0.62	4.78	2.77E-04
cellular carbohydrate metabolic process	5.25	1.80	2.91	3.27E-04
carbohydrate metabolic process	7.76	3.37	2.30	3.53E-04
response to radiation	5.48	1.98	2.77	4.20E-04
oligosaccharide metabolic process	1.83	0.23	8.00	4.87E-04
cell growth	3.65	1.01	3.63	5.46E-04
response to light stimulus	5.25	1.91	2.75	6.64E-04
cellular component biogenesis	2.97	0.71	4.21	7.28E-04
cellular component biogenesis at cellular level	2.97	0.71	4.21	7.28E-04
cellular carbohydrate biosynthetic process	2.97	0.71	4.16	7.90E-04
embryo development ending in seed dormancy	4.34	1.42	3.05	8.05E-04
growth	3.88	1.18	3.30	8.18E-04
cellulose biosynthetic process	1.37	0.12	11.03	9.09E-04
intracellular transport	4.57	1.57	2.90	9.09E-04
response to gravity	1.60	0.19	8.31	1.00E-03
unidimensional cell growth	2.97	0.76	3.92	1.21E-03
cellulose metabolic process	1.37	0.14	10.06	1.29E-03
developmental growth	3.20	0.88	3.64	1.29E-03
establishment of localization in cell	4.79	1.79	2.68	1.50E-03
DNA unwinding involved in replication	0.91	0.04	22.80	1.58E-03
pollen germination	1.37	0.14	9.50	1.60E-03
protein complex assembly	2.28	0.48	4.71	1.88E-03
cellular process	42.69	33.96	1.26	1.98E-03
root development	2.51	0.59	4.24	1.98E-03
developmental growth involved in morphogenesis	2.97	0.83	3.60	2.24E-03
anatomical structure development	5.48	2.31	2.37	2.48E-03
establishment of protein localization	4.11	1.49	2.76	2.86E-03

Biological process GO term	**Observed frequency (%)**	**Expected frequency (%)**	**Ratio**	***p*-value**
gravitropism	1.37	0.17	8.14	2.86E-03
intracellular protein transport	3.42	1.09	3.13	2.86E-03
protein transport	4.11	1.49	2.76	2.86E-03
cellular protein complex assembly	2.05	0.42	4.84	2.90E-03
DNA duplex unwinding	0.91	0.06	16.28	3.42E-03
DNA geometric change	0.91	0.06	16.28	3.42E-03
positive regulation of post-embryonic development	1.37	0.18	7.77	3.42E-03
vacuole organization	0.91	0.06	16.28	3.42E-03
protein complex subunit organization	2.28	0.55	4.16	3.73E-03
pattern specification process	2.05	0.46	4.46	4.62E-03
trichome morphogenesis	1.14	0.12	9.50	4.67E-03
response to abiotic stimulus	9.59	5.45	1.76	5.50E-03
cell tip growth	1.83	0.38	4.85	5.61E-03
organ development	2.97	0.95	3.13	6.10E-03
positive regulation of developmental process	1.37	0.20	6.70	6.10E-03
tropism	1.37	0.21	6.58	6.61E-03
Golgi organization	0.68	0.03	24.42	7.54E-03
pollen tube growth	1.60	0.30	5.25	7.74E-03

Appendix C

Software

C.1 Peptide array image analysis software

Software with a graphical user interface (GUI) was used to analyse SIM peptide array images. The GUI allows a grid to be specified over the desired image and the intensity values under the grid can then be calculated. An example of the interface is shown in Figure C.1. The main GUI code is stored in the file `Array_Tool.m` and depends on the functions shown in Table C.1.

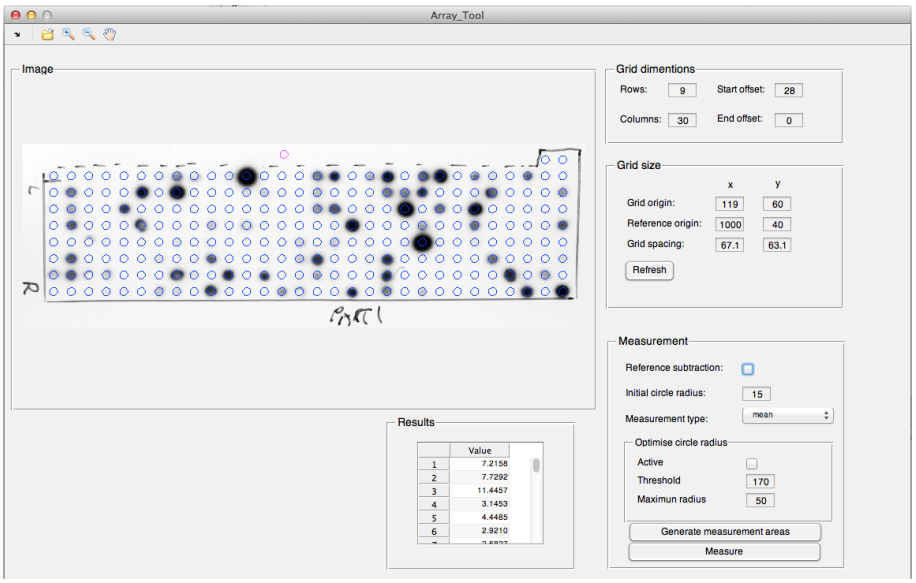


Figure C.1: Peptide array analysis GUI interface.

Function name	Purpose
`calcGrid.m`	Calculates the coordinates of grid points based on grid size, spacing and offset.
`discCalc.m`	Calculates pixel coordinates for all pixels within a circle with specified radius and centre-point, then calculates either the mean, median or sum of pixel values within the disc.
`drawCircles.m`	Draws circles onto peptide array image based on results radii and circle midpoint coordinates.
`drawDots.m`	Draws dots onto peptide array image based on grid centre-point coordinates.
`drawFigure.m`	Redraws peptide array image, removing anything that was drawn over image previously.
`generateResults.m`	Iterates the `discCalc.m` over all points in the grid and outputs the data as a table.
`getmidpointcircle.m`	Calculate discretised pixel coordinates around a centre point using the get mid point circle algorithm. Source code from Tinevez (2013), retrieved from MATLAB Central File Exchange.
`Array_Tool.m`	Main function for generating GUI, Calls other functions to perform actions.

Table C.1: Function files from the peptide image analysis software tool.

C.1.1 `CalcGrid.m` function

```
function calcGrid(hObject)
%Draws mid-points of circles

%Variables
handles = guidata(gcbo);
xOr     = str2num(get(handles.gridX, 'string'));
yOr     = str2num(get(handles.gridY, 'string'));
xN      = str2num(get(handles.nCols, 'string'));
yN      = str2num(get(handles.nRows, 'string'));
xSp     = str2num(get(handles.cellSizeX, 'string'));
ySp     = str2num(get(handles.cellSizeY, 'string'));
startOf = str2num(get(handles.startOff, 'string'));
endOf   = str2num(get(handles.endOff, 'string'));
refX    = str2num(get(handles.refX, 'string'));
refY    = str2num(get(handles.refY, 'string'));

%Generate grid
[X Y] = meshgrid(1:xN, 1:yN);
X = X';
Y = Y';
X = X(:)*xSp + xOr - xSp;
Y = Y(:)*ySp + yOr - ySp;

%Remove offset
X = X((1+startOf):(length(X)-endOf));
Y = Y((1+startOf):(length(Y)-endOf));

%Plot points on graph
%hold on
%plot(X, Y, '+')
%plot(refX, refY, '+', 'color', 'magenta')
%hold off

%Save new data
handles.X = X;
handles.Y = Y;
guidata(hObject, handles);
```

C.1.2 `discCalc.m` function

```
function [result n] = discCalc(image, x, y, r, mode)
%DISCCALC calculates the sum, mean or median of pixels in a disc
%   mode can be:   'mean'   result = mean
%                  'median' result = median
%                  'sum'    result = sum
%                  'raw'    result = vector of values
%   n in output is the number of points that satisfy the
%     disc criteria

[columnsInImage rowsInImage] = meshgrid(1:size(image,2), 1:size(image,1));
% Next create the circle in the image.
circlePixels = logical((rowsInImage - y).^2 ...
    + (columnsInImage - x).^2 <= r.^2);
% circlePixels is a 2D "logical" array.
% Now, display it.
n = sum(sum(circlePixels));

if strcmp(mode, 'mean'),
    result = mean(image(circlePixels));
elseif strcmp(mode, 'median'),
    result = median(image(circlePixels));
elseif strcmp(mode, 'sum'),
    result = sum(image(circlePixels));
elseif strcmp(mode, 'raw'),
    result = image(circlePixels);
else
```

```
    error('Invalid type specified in input paramters')
end

end
```

C.1.3 drawCircles.m function

```
function drawCircles(hObject)
calcGrid(hObject)
%Draws circles
%Variables
handles      = guidata(gcbo);
initRadius   = str2double(get(handles.initCircSize, 'string'));
threshold    = str2double(get(handles.threshold, 'string'));
limit        = str2double(get(handles.limit, 'string'));
n            = length(handles.X);
sizeOptimise = get(handles.sizeOptimise, 'value');
X            = round(handles.X);
Y            = round(handles.Y);
invImg       = imcomplement(handles.img);
refX         = str2num(get(handles.refX, 'string'));
refY         = str2num(get(handles.refY, 'string'));

%initialise radii variable with minimum circle size
radii = ones(n, 1) * initRadius;

%Make circles
switch sizeOptimise
    case 0
        %No size optimisation
        for i = 1:n,
            hold on
            [x y] = getmidpointcircle(X(i), Y(i), radii(i));
            plot(x, y)
            hold off
        end
    case 1
        %With size optimisation
        for i = 1:n,
            [x y] = getmidpointcircle(X(i), Y(i), radii(i));
            while mean(invImg(sub2ind(size(invImg),y,x))) > ...
                    threshold && radii(i) < limit,
                radii(i) = radii(i) + 1;
                [x y] = getmidpointcircle(X(i), Y(i), radii(i));
            end
            hold on
            plot(x, y)
            hold off
        end
    otherwise
        error('Critical error')
end

[x y] = getmidpointcircle(refX, refY, initRadius);
hold on
plot(x, y, 'color', 'magenta')
hold off

%Save data
handles.radii = radii;
guidata(hObject, handles)
```

C.1.4 drawDots.m function

```
function drawDots(hObject)
%Draws mid-points of circles
```

```
calcGrid(hObject)

%Variables
handles = guidata(gcbo);
refX    = str2num(get(handles.refX, 'string'));
refY    = str2num(get(handles.refY, 'string'));

%Plot points on graph
hold on
plot(handles.X, handles.Y, '+')
plot(refX, refY, '+', 'color', 'magenta')
hold off
```

C.1.5 drawFigure.m function

```
function drawFigure
handles = guidata(gcbo);
imshow(handles.img);
```

C.1.6 generateResults.m function

```
function generateResults(hObject)
%Variables
handles     = guidata(gcbo);
invImg      = double(imcomplement(handles.img));
X           = round(handles.X);
Y           = round(handles.Y);
radii       = handles.radii;
temp        = get(handles.measType,{'String','Value'});
measureType = temp{1}{temp{2}};
n           = length(X);
output      = zeros(n,1);
refSub      = get(handles.refSub, 'value');
refX        = str2num(get(handles.refX, 'string'));
refY        = str2num(get(handles.refY, 'string'));
initRadius  = str2double(get(handles.initCircSize, 'string'));

switch refSub
    case 0
        for i = 1:n,
            output(i) = discCalc(invImg, X(i), Y(i), radii(i), measureType);
        end
    case 1
        switch measureType
            case 'mean'
                background = discCalc(invImg, refX, refY, ...
                    initRadius, measureType);
                for i = 1:n,
                    output(i) = discCalc(invImg, X(i), Y(i), ...
                        radii(i), measureType) - background;
                end
            case 'median'
                background = discCalc(invImg, refX, refY, ...
                    initRadius, measureType);
                for i = 1:n,
                    output(i) = discCalc(invImg, X(i), Y(i), ...
                        radii(i), measureType) - background;
                end
            case 'sum'
                [background bPixls] = discCalc(invImg, refX, refY, ...
                    initRadius, measureType);
                background = background/bPixls;
                for i = 1:n,
                    [output(i) pixls] = discCalc(invImg, X(i), Y(i), ...
                        radii(i), measureType);
                    output(i) = output(i) - background*pixls;
```

```
                end
            otherwise
                error('Critical error')
        end
    otherwise
        error('Critical error')
end



set(handles.resultsTable, 'Data', output)
guidata(hObject, handles);
```

C.1.7 `getmidpointcircle.m` function

The following function was written by Tinevez (2013) and was retrieved from the MATLAB Central File Exchange.

```
function [xc, yc] = getmidpointcircle(x0, y0, radius)
%% GETMIDPOINTCIRCLE return the x,y pixel coordinates of a circle
%
% [x y] = GETMIDPOINTCIRCLE(x0, y0, radius) returns the pixel coordinates
% of the circle centered at pixel position [x0 y0] and of the given integer
% radius. The mid-point circle algorithm is used for computation
% (http://en.wikipedia.org/wiki/Midpoint_circle_algorithm).
%
% This function is aimed at image processing applications, where the
% integer pixel coordinates matter, and for which one pixel cannot be
% missed or duplicated. In that view, using rounded trigonometric
% coordinates generated using cosine calls are inadequate. The mid-point
% circle algorithm is the answer.
%
% Accent is made on performance. We compute in advance the number of point
% that will be generated by the algorithm, to pre-allocate the coordinates
% arrays. I have tried to do this using a MATLAB class implementing the
% iterator pattern, to avoid computing the number of points in advance and
% still be able to iterate over circle points. However, it turned out that
% repeated function calls is extremely expansive, and the class version of
% this function is approximately 1000 times slower. With this function, you
% can get the pixel coordinates of a circle of radius 1000 in 0.16 ms, and
% this time will scale linearly with increasing radius (e.g. it takes
% 0.16 s for a radius of 1 million).
%
% Also, this functions ensure that sorted coordinates are returned. The
% mid-point algorithm normally generates a point for the 8 circles octants
% in one iteration. If they are put in an array in that order, the [x y]
% points will jump from one octant to another. Here, we ensure that they
% are returned in order, starting from the top point, and going clockwise.
%
% EXAMPLE
%
% n_circles = 20;
% color_length = 100;
% image_size = 128;
% max_radius = 20;
%
% I = zeros(image_size, image_size, 3, 'uint8');
% colors = hsv(color_length);
%
% for i = 1 : n_circles
%
%     x0 = round( image_size * rand);
%     y0 = round( image_size * rand);
%     radius = round( max_radius * rand );
%
%     [x y] = getmidpointcircle(x0, y0, radius);
%
%     index = 1 ;
%     for j = 1 : numel(x)
%         xp = x(j);
```

```
53 %         yp = y(j);
54 %
55 %         if ( xp < 1 || yp < 1 || xp > image_size || yp > image_size )
56 %             continue
57 %         end
58 %         I(xp, yp, :) = round( 255 * colors(index, :) );
59 %         index = index + 1;
60 %         if index > color_length
61 %             index = 1;
62 %         end
63 %     end
64 %
65 % end
66 %
67 % imshow(I, []);
68 %
69 %
70 % Jean-Yves Tinevez <jeanyves.tinevez@gmail.com> - Nov 2011 - Feb 2012
71
72     % Compute first the number of points
73     octant_size = floor((sqrt(2)*(radius-1)+4)/2);
74     n_points = 8 * octant_size;
75
76     % Iterate a second time, and this time retrieve coordinates.
77     % We "zig-zag" through indices, so that we reconstruct a continuous
78     % set of of x,y coordinates, starting from the top of the circle.
79
80     xc = NaN(n_points, 1);
81     yc = NaN(n_points, 1);
82
83     x = 0;
84     y = radius;
85     f = 1 - radius;
86     dx = 1;
87     dy = - 2 * radius;
88
89     % Store
90
91     % 1 octant
92     xc(1) = x0 + x;
93     yc(1) = y0 + y;
94
95     % 2nd octant
96     xc(8 * octant_size) = x0 - x;
97     yc(8 * octant_size) = y0 + y;
98
99     % 3rd octant
100    xc(4 * octant_size) = x0 + x;
101    yc(4 * octant_size) = y0 - y;
102
103    % 4th octant
104    xc(4 * octant_size + 1) = x0 - x;
105    yc(4 * octant_size + 1) = y0 - y;
106
107    % 5th octant
108    xc(2 * octant_size) = x0 + y;
109    yc(2 * octant_size) = y0 + x;
110
111    % 6th octant
112    xc(6 * octant_size + 1) = x0 - y;
113    yc(6 * octant_size + 1) = y0 + x;
114
115    % 7th octant
116    xc(2 * octant_size + 1) = x0 + y;
117    yc(2 * octant_size + 1) = y0 - x;
118
119    % 8th octant
120    xc(6 * octant_size) = x0 - y;
121    yc(6 * octant_size) = y0 - x;
122
123
124    for i = 2 : n_points/8
125
```

```
126             % We update x & y
127             if f > 0
128                 y = y - 1;
129                 dy = dy + 2;
130                 f = f + dy;
131             end
132             x = x + 1;
133             dx = dx + 2;
134             f = f + dx;
135 
136             % 1 octant
137             xc(i) = x0 + x;
138             yc(i) = y0 + y;
139 
140             % 2nd octant
141             xc(8 * octant_size - i + 1) = x0 - x;
142             yc(8 * octant_size - i + 1) = y0 + y;
143 
144             % 3rd octant
145             xc(4 * octant_size - i + 1) = x0 + x;
146             yc(4 * octant_size - i + 1) = y0 - y;
147 
148             % 4th octant
149             xc(4 * octant_size + i) = x0 - x;
150             yc(4 * octant_size + i) = y0 - y;
151 
152             % 5th octant
153             xc(2 * octant_size - i + 1) = x0 + y;
154             yc(2 * octant_size - i + 1) = y0 + x;
155 
156             % 6th octant
157             xc(6 * octant_size + i) = x0 - y;
158             yc(6 * octant_size + i) = y0 + x;
159 
160             % 7th octant
161             xc(2 * octant_size + i) = x0 + y;
162             yc(2 * octant_size + i) = y0 - x;
163 
164             % 8th octant
165             xc(6 * octant_size - i + 1) = x0 - y;
166             yc(6 * octant_size - i + 1) = y0 - x;
167 
168         end
169 
170 end
```

C.1.8 `Array_Tool.m` GUI function

```
1  function varargout = Array_Tool(varargin)
2  % ARRAY_TOOL MATLAB code for Array_Tool.fig
3  %      ARRAY_TOOL, by itself, creates a new ARRAY_TOOL or raises the existing
4  %      singleton*.
5  %
6  %      H = ARRAY_TOOL returns the handle to a new ARRAY_TOOL or the handle to
7  %      the existing singleton*.
8  %
9  %      ARRAY_TOOL('CALLBACK',hObject,eventData,handles,...) calls the local
10 %      function named CALLBACK in ARRAY_TOOL.M with the given input arguments.
11 %
12 %      ARRAY_TOOL('Property','Value',...) creates a new ARRAY_TOOL or raises the
13 %      existing singleton*.  Starting from the left, property value pairs are
14 %      applied to the GUI before Array_Tool_OpeningFcn gets called.  An
15 %      unrecognized property name or invalid value makes property application
16 %      stop.  All inputs are passed to Array_Tool_OpeningFcn via varargin.
17 %
18 %      *See GUI Options on GUIDE's Tools menu.  Choose "GUI allows only one
19 %      instance to run (singleton)".
20 %
21 % See also: GUIDE, GUIDATA, GUIHANDLES
22 
```

```
% Edit the above text to modify the response to help Array_Tool

% Last Modified by GUIDE v2.5 17-Mar-2014 01:45:50

% Begin initialization code - DO NOT EDIT
gui_Singleton = 1;
gui_State = struct('gui_Name',       mfilename, ...
                   'gui_Singleton',  gui_Singleton, ...
                   'gui_OpeningFcn', @Array_Tool_OpeningFcn, ...
                   'gui_OutputFcn',  @Array_Tool_OutputFcn, ...
                   'gui_LayoutFcn',  [] , ...
                   'gui_Callback',   []);
if nargin && ischar(varargin{1})
    gui_State.gui_Callback = str2func(varargin{1});
end

if nargout
    [varargout{1:nargout}] = gui_mainfcn(gui_State, varargin{:});
else
    gui_mainfcn(gui_State, varargin{:});
end
% End initialization code - DO NOT EDIT


%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%% Update graph with centre points after text change                        %%%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

function nRows_Callback(hObject, eventdata, handles) %#ok<*DEFNU,*INUSD>
drawFigure
drawDots(hObject)

function startOff_Callback(hObject, eventdata, handles)
drawFigure
drawDots(hObject)

function nCols_Callback(hObject, eventdata, handles)
drawDots(hObject)

function endOff_Callback(hObject, eventdata, handles)
drawFigure
drawDots(hObject)


function gridX_Callback(hObject, eventdata, handles)
drawFigure
drawDots(hObject)

function refX_Callback(hObject, eventdata, handles)
drawFigure
drawDots(hObject)

function cellSizeX_Callback(hObject, eventdata, handles)
drawFigure
drawDots(hObject)

function gridY_Callback(hObject, eventdata, handles)
drawFigure
drawDots(hObject)

function refY_Callback(hObject, eventdata, handles)
drawFigure
drawDots(hObject)


function cellSizeY_Callback(hObject, eventdata, handles)
drawFigure
drawDots(hObject)

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%% Other callbacks                                                          %%%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

```

```
% Draw cirles of measruing area over dots
function genCirc_Callback(hObject, eventdata, handles)
drawFigure
drawCircles(hObject)

% Draw points after button click
function test_Callback(hObject, eventdata, handles)
drawFigure
drawDots(hObject)

% Generates intensity results on button click
function pushbutton3_Callback(hObject, eventdata, handles)
generateResults(hObject)

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%% User defined functions                                                   %%%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

% Load new image
function opFile_ClickedCallback(hObject, eventdata, handles)
handles     = guidata(gcbo);

[filename path] = uigetfile(...
    {'*.jpg; *.jpeg; *.tif; *.tif; *.raw; *.png; *.bmp; *.txt', ...
    'All Image Files'; ...
    '*.*', 'All Files'}, 'Open image file', handles.currentPath);

if filename ~= 0,
    handles.img = imread([path filename]);
    imshow(handles.img);
end

%Save fold location
handles.currentPath = path;

% Update handles structure
guidata(hObject, handles);


% --------------------------------------------------------------------
%  END OF USER DEFINED FUNCTIONS
%    The following code is machine generated and is required for GUI
% --------------------------------------------------------------------

% --- Executes just before Array_Tool is made visible.
function Array_Tool_OpeningFcn(hObject, eventdata, handles) %#ok<*INUSL>
handles.output = hObject;
handles.img = imread('sampleImage.jpg');
imshow(handles.img)
handles.currentPath = [];
guidata(hObject, handles);

% --- Outputs from this function are returned to the command line.
function varargout = Array_Tool_OutputFcn(hObject, eventdata, handles)
varargout{1} = handles.output;

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%% Executes during object creation, after setting all properties.      %%%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
function edit2_CreateFcn(hObject, eventdata, handles)
if ispc && isequal(get(hObject,'BackgroundColor'), ...
        get(0,'defaultUicontrolBackgroundColor'))
    set(hObject,'BackgroundColor','white');
end

function startOff_CreateFcn(hObject, eventdata, handles)

if ispc && isequal(get(hObject,'BackgroundColor'), ...
        get(0,'defaultUicontrolBackgroundColor'))
    set(hObject,'BackgroundColor','white');
end

function nCols_CreateFcn(hObject, eventdata, handles)
```

```
169 if ispc && isequal(get(hObject,'BackgroundColor'), ...
170         get(0,'defaultUicontrolBackgroundColor'))
171     set(hObject,'BackgroundColor','white');
172 end
173
174 function endOff_CreateFcn(hObject, eventdata, handles)
175
176 if ispc && isequal(get(hObject,'BackgroundColor'), ...
177         get(0,'defaultUicontrolBackgroundColor'))
178     set(hObject,'BackgroundColor','white');
179 end
180
181 function gridX_CreateFcn(hObject, eventdata, handles)
182
183 if ispc && isequal(get(hObject,'BackgroundColor'), ...
184         get(0,'defaultUicontrolBackgroundColor'))
185     set(hObject,'BackgroundColor','white');
186 end
187
188 function refX_CreateFcn(hObject, eventdata, handles)
189 if ispc && isequal(get(hObject,'BackgroundColor'), ...
190         get(0,'defaultUicontrolBackgroundColor'))
191     set(hObject,'BackgroundColor','white');
192 end
193
194 function cellSizeX_CreateFcn(hObject, eventdata, handles)
195 if ispc && isequal(get(hObject,'BackgroundColor'), ...
196         get(0,'defaultUicontrolBackgroundColor'))
197     set(hObject,'BackgroundColor','white');
198 end
199
200 function gridY_CreateFcn(hObject, eventdata, handles)
201
202 if ispc && isequal(get(hObject,'BackgroundColor'), ...
203         get(0,'defaultUicontrolBackgroundColor'))
204     set(hObject,'BackgroundColor','white');
205 end
206
207 function refY_CreateFcn(hObject, eventdata, handles)
208
209 if ispc && isequal(get(hObject,'BackgroundColor'), ...
210         get(0,'defaultUicontrolBackgroundColor'))
211     set(hObject,'BackgroundColor','white');
212 end
213
214 function cellSizeY_CreateFcn(hObject, eventdata, handles)
215 if ispc && isequal(get(hObject,'BackgroundColor'), ...
216         get(0,'defaultUicontrolBackgroundColor'))
217     set(hObject,'BackgroundColor','white');
218 end
219
220 function initCircSize_CreateFcn(hObject, eventdata, handles)
221 if ispc && isequal(get(hObject,'BackgroundColor'), ...
222         get(0,'defaultUicontrolBackgroundColor'))
223     set(hObject,'BackgroundColor','white');
224 end
225
226 function measType_CreateFcn(hObject, eventdata, handles)
227 if ispc && isequal(get(hObject,'BackgroundColor'), ...
228         get(0,'defaultUicontrolBackgroundColor'))
229     set(hObject,'BackgroundColor','white');
230 end
231
232 function limit_CreateFcn(hObject, eventdata, ~)
233 if ispc && isequal(get(hObject,'BackgroundColor'), ...
234         get(0,'defaultUicontrolBackgroundColor'))
235     set(hObject,'BackgroundColor','white');
236 end
237
238 function threshold_CreateFcn(hObject, eventdata, handles)
239 if ispc && isequal(get(hObject,'BackgroundColor'), ...
240         get(0,'defaultUicontrolBackgroundColor'))
241     set(hObject,'BackgroundColor','white');
```

```
end

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%% Unused callback functions                                           %%%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
function togglebutton1_Callback(hObject, eventdata, handles)
function nRows_CreateFcn(hObject, eventdata, handles)
function mainGraph_CreateFcn(hObject, eventdata, handles)
function refSub_Callback(hObject, eventdata, handles)
function nRows_ButtonDownFcn(hObject, eventdata, handles)
function initCircSize_Callback(hObject, eventdata, handles)
function measType_Callback(hObject, eventdata, handles)
function sizeOptimise_Callback(hObject, eventdata, handles)
function limit_Callback(hObject, eventdata, handles)
function threshold_Callback(hObject, eventdata, handles)
```

C.2 Sequence analysis functions

C.2.1 Sequence similarity

The following C source code is for the sequence similarity shared library used by R to calculate sequence similarity using the sum of pairs method.

```
/*  C code for an object to be used by R for efficiently calculating
 *  protein conservation using a substation matrix to score amino acid pairs.
 *
 *  Object accepts pointers to R objects to be operated on.
 *
 *  The R objects are 1 dimensional arrays that were coerced from
 *  2 dimensional R arrays. The arrays are kept as 1 dimensional arrays and
 *  and 1 dimensional indices are calculated from 2 dimensional indices.
 */

#include <stdio.h>
#include <stdlib.h>

/*  Recursive function to calculate binomial coefficients
 *   This function is used to calculate the number of  of amino-
 *   acid substitution pairs. If there are N sequences, there
 *   (N choose 2) pair substations.
 */
int nchoosek(int n, int k)
{
    if (k == 0) return 1;
    if (n == 0) return 0;
    return(nchoosek(n-1, k-1) + nchoosek(n-1, k));
}

void HSDS_conservation_C(int *nLetters, int *nSequences, int *nSmoothing,
   char **chAlignment, double *dOutput, double *dSCORE,
   char **chAlphabet, int *nAlphabetsize)
 {
    /* Convert alignment matrix into an index matrix for dSCORE.
     *  This will allow using alignment values to directly index the
     *  substitution matrix, rather than having to look up values.
     */

    // Allocate memory for the index matrix, nAlignment
    int **nAlignment;
    nAlignment = malloc(nSequences[0] * sizeof(int *));
    for (int i = 0; i < nSequences[0]; i++)
    {
       nAlignment[i] = malloc(nLetters[0] * sizeof(int));
    }

    // Convert amino acids into their respective integer index values and
    //  populate the nAlignment matrix with these values
    for (int i = 0; i < nSequences[0]; i++)
    {
       for (int j = 0; j < nLetters[0]; j++)
       {
          for (int k = 0; k < nAlphabetsize[0]; k++)
          {
             if (chAlignment[(i * nLetters[0]) + j][0] == chAlphabet[k][0])
             {
                nAlignment[i][j] = k;
             }
          }
       }
    }


    /*
     * Calculate conservation
```

```
62        */
63
64      // Calculate the number pair substitutions, N, to, which will be used in
65      // the mean calculation.
66      double fDenominator = (double)nchoosek(nSequences[0], 2);
67
68      // For each position in the alignment, calculate the sum of possible
69      //  pair substitution then divide by N.
70      for (int i = 0; i < nLetters[0]; i++)
71      {
72         for (int j = 0; j < (nSequences[0] - 1); j++)
73         {
74            for (int k = j + 1; k < nSequences[0]; k++)
75            {
76               dOutput[i] += dSCORE[ (nAlignment[j][i] * nAlphabetsize[0]) +
77                  nAlignment[k][i] ];
78            }
79         }
80         dOutput[i] /= fDenominator;
81      }
82
83
84      // Smooth data, averages each position by a window nSmoothing characters
85      //  upstream and downstream of each position
86      if (nSmoothing[0] > 0)
87      {
88         double *dTemp = malloc(nLetters[0] * sizeof(double));
89         for (int i = 0; i < nLetters[0]; i++)
90         {
91            dTemp[i] = dOutput[i];
92         }
93
94         int a, b;
95         double dSum;
96         for (int i = 0; i < nLetters[0]; i++)
97         {
98            a = i - nSmoothing[0];
99            b = i + nSmoothing[0];
100           dSum = 0.0;
101           if (a < 0)
102              a = 0;
103           if (b >= nLetters[0])
104              b = nLetters[0] - 1;
105           for (int j = a; j <= b; j++)
106           {
107              dSum += dTemp[j];
108           }
109           dOutput[i] = dSum / (double)(b - a + 1);
110        }
111     }
112  }
```

R wrapper function for the C shared library function:

```
1  HSDS_conservation <- function(alignment, nSmoothing = 0, useCompiled = TRUE,
2                                matrix_dir = "Simplant/Data/",
3                                c_dir = "Simplant/Executables/"){
4    # The data file for the score matrix needs to present in matrix_dir.
5    # To use the C compiled version function which is much faster,
6    #   either the 'HSDS_conservation_C.so' or 'HSDS_conservation_C.c'
7    #   needs to be present in the c_dir directory.
8    if (is.null(alignment)) return(1)
9
10   # Check whether the HSDS matrix is loaded, if not
11   #   try find it and load it.
12   #HSDS <- NULL
13   if(!exists("HSDS")){
14     if (file.exists("Data/HSDS_matrix.rds")){
15       HSDS <- readRDS("Data/HSDS_matrix.rds")
16     }else if(file.exists(paste0(matrix_dir, "HSDS_matrix.rds"))){
```

```
17        HSDS <- readRDS(paste0(matrix_dir, "HSDS_matrix.rds"))
18      }else{
19        stop("Can find HSDS matrix file. Check the 'matrix_dir' folder name.")
20      }
21    }
22
23    # Variables
24    nLetters     <- ncol(alignment$ali)
25    nSequences   <- nrow(alignment$ali)
26    output       <- rep(0, nLetters)
27
28    # If there is only 1 sequence present, return output of zeros here
29    if(nSequences == 1){
30      return(output)
31    }
32
33    # If using compiled, check if possible,
34    #   if standard library file can't be
35    #   found, a warning is issued and function
36    #   jumps to R code instead.
37    if(useCompiled == TRUE){
38      # First check if .so is loaded
39      if (!is.loaded("HSDS_conservation_C")){
40
41        # If not check for the .so file then load it
42        if(file.exists("Executables/HSDS_conservation_C.so")){
43          dyn.load("Executables/HSDS_conservation_C.so")
44        }else if(file.exists(paste0(c_dir, "HSDS_conservation_C.so"))){
45          dyn.load(paste0(c_dir, "HSDS_conservation_C.so"))
46
47          # If there is no .so file, check for the .c file
48          #  and try to compile it and then load .so file
49        }else if(file.exists("Executables/HSDS_conservation_C.c")){
50          system("R CMD SHLIB Executables/HSDS_conservation_C.c")
51          dyn.load("Executables/HSDS_conservation_C.so")
52        }else if(file.exists(paste0(c_dir, "HSDS_conservation_C.so"))){
53          system(paste0("R CMD SHLIB ", c_dir, "HSDS_conservation_C.c"))
54          dyn.load(paste0(c_dir, "HSDS_conservation_C.c"))
55
56          # If the .c file cant be found, issue a warning and use
57          #  use R coded part of the function instead
58        }else{
59          warning(paste("Can't find shared object file, using slower R",
60                        "code function instead.",
61                        "Check that the directory is correct and the",
62                        "'HSDS_conservation_C.c' file is present."))
63          useCompiled <- FALSE
64        }
65      }
66    }
67
68    if(useCompiled == TRUE){
69      ### Do C version
70      #nLetters     <- ncol(alignment$ali)
71      #nSequences   <- nrow(alignment$ali)
72      #output       <- rep(0, nLetters)
73
74      output <- .C("HSDS_conservation_C",
75                   nLetters      = as.integer(ncol(alignment$ali)),
76                   nSequences    = as.integer(nrow(alignment$ali)),
77                   nSmoothing    = as.integer(nSmoothing),
78                   chAlignment   = as.vector(t(alignment$ali)),
79                   dOutput       = as.numeric(rep(0, ncol(alignment$ali))),
80                   dSCORE        = as.vector(HSDS),
81                   chAlphabet    = rownames(HSDS),
82                   nAlphabetsize = as.integer(length(rownames(HSDS)))
83      )$dOutput
84
85    }else{
86      ### Do R version
87      # Conservation scoring
88      n <- choose(nSequences, 2)
89      for (i in 1:nLetters){
```

```
90        for(j in 1:(nSequences - 1)){
91          for(k in (j+1):nSequences){
92            output[i] <- output[i] +
93              HSDS[alignment$ali[j, i], alignment$ali[k, i]]
94          }
95        }
96        output[i] <- output[i]/n
97      }
98
99      #Smoothing
100     if(nSmoothing > 0){
101       temp <- output
102       for(i in 1:nLetters){
103         a <- i - nSmoothing
104         b <- i + nSmoothing
105         if (a < 1) a <- 1
106         if (b > nLetters) b <- nLetters
107         output[i] <- mean(temp[a:b])
108       }
109     }
110   }
111
112   return(output)
113 }
114
115 # Complile function to bytecode
116 HSDS_conservation <- cmpfun(HSDS_conservation)
```

C.2.2 Preference logo

The following function is written in R.

```
1  preferenceLogo <- function(alignment, background = "arabidopsis",
2                             plot = TRUE, output = FALSE, title="",
3                             MCcorrection = "none", font = "Arial",
4                             fontsize = 0.7){
5    ### Plot a preference logo
6    #      Plot is based on a Berry style logo as an alternative to
7    #       sequence logos. This logo displays amino acid frequencies
8    #       relative to a background frequency. This is in contrast
9    #       sequence logos which display absolute frequencies scaled
10   #       by information content at each position.
11   #       The purpose of the preference logo is compare the
12   #       distribution of two sets of sequences. The background
13   #       set frequencies are subtracted from the test set
14   #       ("alignment" variable) and amino acid comparisons with
15   #       too small data amount are excluded using a Poisson test
16   #       where the p value is greater than 0.05. This solves issues
17   #       when comparing datasets with a small number of observations.
18   #
19   #     Parameters:
20   #      alignment: Test set of aligned, non gapped sequences. This
21   #                 must be a matrix of characters corresponding to
22   #                 the single letter amino alphabet. Matrix rows
23   #                 represent individual sequences and columns
24   #                 represent sequence posisitions.
25   #      background: This can either be a matrix of background
26   #                 sequences or a character string specifing
27   #                 species specific amino acid frequencies. If
28   #                 a matrix is used, the must have the same number
29   #                 of columns (positions) as the alignment matrix.
30   #                 Current strings for default frequencies are
31   #                 c("arabidopsis", "human").
32   #      MCRcorrection: A multiple comparison correction can be applied
33   #                 to the p value from the Poisson test. The defualt is
34   #                 "none", the methods are c("holm", "hochberg",
35   #                 "hommel", "bonferroni", "BH", "BY", "fdr"). See
36   #                 the help topic on p.adjust from the {stats} package
```

```
#                 for futher explanation

# Required packages
require("rms.gof") # for the Poisson test

# Single letter amino acid alphabet and corresponding
#  hydrophobicity colour scheme
alphabet <- c("I", "V", "L", "F", "C",
              "M", "A", "G", "T", "S",
              "W", "Y", "P", "H", "N",
              "D", "E", "Q", "K", "R")
alphCols <- c(colorRampPalette(c(rgb(.7,.3,0,.6),rgb(.7,.7,0,.6)))(7),
              colorRampPalette(c(rgb(0,0.5,.8,.6),rgb(0,0,.8,.6)))(13))

# Amino acid frequency tables for plant and human
AtFreqs <- matrix(c(3468,4417,6079,2756,1329,
                    1745,3986,4116,3328,5731,
                    823,1912,3259,1504,2822,
                    3471,4210,2216,4312,3516)/65000,
                  1, 20, dimnames = list("", alphabet))
HsFreqs <- matrix(c(5.6, 6.6, 9.1, 3.9, 1.9,
                    2.2, 7.8, 7.2, 5.9, 6.8,
                    1.4, 3.2, 5.2, 2.3, 5.3,
                    6.3, 4.3, 4.2, 5.9, 5.1)/100,
                  1, 20, dimnames = list("", alphabet))
#zeroFreqs

# Calculate variables
nSequences <- dim(alignment)[1]
nPositions <- dim(alignment)[2]
nLetters   <- length(alphabet)

# countTable function takes an input sequence matrix and returns the counts
#   for each amino acid at each position. The returned data is in a matrix
countTable <- function(fastaMatrix){
  output <- matrix(-1, nPositions, nLetters,
                   dimnames = list(NULL, alphabet))
  for(i in 1:nPositions){
    for(j in 1:nLetters){
      output[i,j] <- sum(fastaMatrix[,i]  == alphabet[j])
    }
  }
  return(output)
}

# Make baground counts table
#  The background counts table is scaled to have
#  so that the column sums are the same as the number of sequences in
#  the input alignment matrix, this is required for the later relative
#  counts calculation.
if (background[1] == "arabidopsis"){
  background <- matrix(0, nPositions, 20,
                       dimnames = list(rep("", nPositions),alphabet))
  AtFreqs <- AtFreqs * nSequences
  for (i in 1:nPositions){
    background[i,] <- AtFreqs
  }
} else if (background[1] == "human"){
  background <- matrix(0, nPositions, 20,
                       dimnames = list(rep("", nPositions),alphabet))
  HsFreqs <- HsFreqs * nSequences
  for (i in 1:nPositions){
    background[i,] <- HsFreqs
  }
}else{
  if (nPositions != dim(background)[2]) error(
    "Wrong backgound string or matrix of wrong size supplied")
  n = dim(background)[1]
  background <- (countTable(background)/n)*nSequences
}

# Calculate frequency differences
alignFreqs <- countTable(alignment)
```

```
110    freqDiff <- (alignFreqs - background)/nSequences
111
112    # Do Poisson test and remove non-significant amino acids
113    sig <- matrix(0, nPositions, nLetters, dimnames = list(NULL, alphabet))
114    for (i in 1:nPositions){
115      for (j in 1:nLetters){
116        p <- poisson.test(alignFreqs[i,j], r = background[i,j])$p.value
117        p <- p.adjust(p, method = MCcorrection, n = nLetters)
118        sig[i,j] <- p
119        if (p > 0.05) freqDiff[i,j] <- NaN
120      }
121    }
122
123    # Remove infinite values
124    freqDiff[freqDiff == Inf | freqDiff == -Inf] <- NaN
125
126    # Calculate plotting parameters
127    ymin <- min(c(freqDiff, -0.1), na.rm = TRUE)
128    ymax <- max(c(freqDiff,  0.5), na.rm = TRUE)
129    xmin <- 1
130    xmax <- dim(freqDiff)[1]
131
132    #Function for sizing letters based on p value
133    sig_size <- function(p){
134      output <- 1/(-p*100-1) + 1
135      return(output)
136    }
137
138    # Draw graph
139    if(plot){
140      par(mar=c(4,2,2,1))
141      plot(c(xmin, xmax), c(ymin, ymax), type = "n", xlab="",
142           ylab="", main = title, family = font,
143           cex.axis = fontsize, las = 1)
144      axis(side = 1, at = seq(xmin, xmax, by = 1),
145           labels = FALSE, tcl = -0.2)
146      grid(ny = NA)
147      abline(c(0,0), c(0,5), col = "gray")
148      for(i in 1:dim(freqDiff)[2]){
149        plotIndex <- !is.nan(freqDiff[,i])
150        x <- (1:xmax)[plotIndex]
151        y <- freqDiff[,i][plotIndex]
152        ps <- sig[,i][plotIndex]
153        points(x, y, pch = alphabet[i], col = alphCols[i],
154               cex = 1.2-sig_size(ps), font = 2)
155      }
156    }
157
158    if (output) return(freqDiff)
159  }
```